# Principe de l’analyse factorielle

Philippe Cibois

Professeur à l’université de Versailles – St-Quentin

Version novembre 2006

1 La représentation géométrique ..................................... 3
2 Comment passer du tableau au graphique .................... 7
3 Les calculs de l'analyse factorielle ............................... 14
4 L'analyse en composantes principales......................... 18
5 L'analyse des correspondances.................................. 28

## Introduction

L’analyse factorielle est une technique statistique aujourd’hui surtout utilisée pour dépouiller des enquêtes : elle permet, quand on dispose d’une population d’individus pour lesquelles on possède de nombreux renseignements concernant les opinions, les pratiques et le statut (sexe, âge, etc.), d’en donner une *représentation géométrique*[1], c'est-à-dire en utilisant un graphique qui permet de voir les rapprochements et les oppositions entre les caractéristiques des individus.

Cette technique est déjà centenaire : elle a été créée en 1904 par le psychologue anglais Charles Spearman (inventeur également du coefficient de corrélation de rang), dans le but de mesurer l’intelligence[2]. Sa technique porte le nom aujourd’hui d’analyse factorielle des psychologues. D’autres techniques d’analyse factorielle seront développées ensuite : *l’analyse en composantes principales*[3] (souvent abrégée en ACP) et une variété de celle-ci *l’analyse factorielle des correspondances* (AFC), crée dans les années 1960 par Jean-Paul Benzécri[4]. Du fait de l’essor de l’informatique, cette dernière technique est devenue une technique standard, intégrée dans les grands logiciels statistiques internationaux (SAS, SPSS).

Le but de ce texte est de donner à toute personne qui le désire les connaissances nécessaires pour comprendre correctement les résultats d’une analyse des correspondances publiés dans des revues de sciences sociales ou dans des feuilles d’information des ministères[5].

Il ne s'agit pas ici de faire la théorie de l’analyse factorielle, ce qui suppose des connaissances mathématiques qui, à mon avis, ne sont pas nécessaires pour comprendre le principe de la méthode. Je prends d’autant plus volontiers cette position que l’exposé du principe de la méthode est largement redevable de

---

[1] Selon l’expression de Henry Rouanet

[2] Charles Spearman, General Intelligence Objectively Determined and Measured, *American Journal of Psychology*, 15 (1904), p.201-292. Disponible à http://psychclassics.yorku.ca/Spearman/

[3] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 (1933) :417-441,498-520.

[4] Benzécri, J.P *et al. L'analyse des données*, Paris, Dunod, 1973, vol. 2 : Correspondances

[5] Regroupés par exemple dans les *Recueil d'études sociales* édités par l'INSEE

VISIT…

LANZAROTE Caliente.COM

l'enseignement à l'EHESS d'un mathématicien, Georges Th. Guilbaud à qui je rends ici hommage. Il a montré dans ses cours et ses écrits que les objets mathématiques utilisés par la statistique étaient précisément des *objets* que l'on pouvait appréhender par le calcul numérique, ce qui sera fait ici. L'expérimentation est un mode d'accès fructueux pour les gens qui ne sont pas statisticien de métier, qui font confiance aux mathématiciens pour qu'ils leur proposent des méthodes valides : quand ceux-ci pensent qu'il faut refaire leur démarche théorique pour utiliser correctement leurs méthodes, ils entrainent ceux qui les suivent dans la voie du découragement. Cependant, ceux qui voudraient approfondir dans une perspective théorique, pourront le faire en utilisant les travaux d'Henry Rouanet[6].

*Vingt ans après*

Ce texte provient du fait que le "Que sais-je ?" paru sous le titre *L'analyse factorielle* en 1983 ne répondait plus à la demande. En effet les attentes des lecteurs ne sont plus les mêmes : quand une nouvelle technique apparait, on cherche à comprendre comment cela fonctionne et on soulève volontiers le couvercle pour démonter l'intérieur. Dans les années 1980, je me souviens avoir dû expliquer comment fonctionnait un ordinateur, mais ces temps sont révolus : on n'éprouve plus ce besoin comme pour le téléphone ou pour un moteur électrique. Pour prendre le vocabulaire de la sociologie des sciences[7], l'ordinateur est utilisé aujourd'hui comme une *boite noire* : on veut n'en connaître que ce qui est utile à un bon usage.

Il en est de même pour les techniques statistiques : en vingt ans d'enseignement régulier de ces techniques, j'ai vu la demande des utilisateurs évoluer, passant d'un désir très fort de savoir comment l'analyse factorielle produisait ses résultats à un objectif différent, comment bien utiliser la méthode. Un nouveau "Que sais-je *?*" *Les techniques d'analyse d'enquête* (à paraitre en 2007) prend acte de cette évolution : la part du principe de la méthode y est réduite pour laisser plus de place à des exemples commentés d'utilisation et à des règles de bonne pratique.

Le présent texte est destiné à ceux qui voudraient cependant ouvrir sérieusement la boite noire de l'analyse factorielle : après une première présentation accessible à tous (pages 3 à 13), ils y trouveront une présentation des calculs qui reprend et détaille les pages correspondantes que premier "Que sais-je ?" désormais non réédité[8].

---

[6] Henry Rouanet et Brigitte Le Roux, *Analyse des données multidimensionnelles*, Paris, Dunod, 1993.

[7] Dominique Vinck, *Sociologie des sciences*, Paris, A. Colin, 1995

[8] Je remercie Bernard Courtebras pour sa relecture attentive de ce texte.

# 1 La représentation géométrique

*Le tableau des carrières*

Partons d'un exemple simplifié d'un tableau de données (appelé dans la suite, *Tableau des carrières*) qui indique ce que deviennent des cadres quand ils changent d'entreprise[9]. En ligne on trouve la position d'origine (préfixée 1) et en colonne la position de destination (préfixée 2). A l'intersection d'une ligne et d'une colonne se trouve le nombre d'individus venant de la position en ligne et s'étant dirigés vers la position en colonne. On a les positions suivantes :

PDG : Président directeur général
DMK : Directeur du marketing
DFI : Directeur financier
DTU : Directeur technique ou d'usine
CBU : Contrôleur budgétaire
DRV : Directeur régional des ventes
IPR : Ingénieur de production
IBE : Ingénieur de bureau d'études
CCO : Cadre comptable
VEN : Acheteur/inspecteur de ventes

| | Destination | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Orig.** | **2PDG** | **2DMK** | **2DFI** | **2DTU** | **2CBU** | **2DRV** | **2IPR** | **2IBE** | **2CCO** | **2VEN** | **Total** |
| **1PDG** | 20 | 3 | 5 | 5 | 0 | 1 | 1 | 0 | 1 | 0 | 36 |
| **1DMK** | 22 | 33 | 0 | 0 | 0 | 9 | 1 | 0 | 1 | 8 | 74 |
| **1DFI** | 10 | 1 | 38 | 0 | 10 | 0 | 0 | 0 | 8 | 1 | 58 |
| **1DTU** | 18 | 0 | 1 | 34 | 1 | 0 | 14 | 9 | 2 | 0 | 79 |
| **1CBU** | 2 | 0 | 12 | 1 | 17 | 2 | 0 | 0 | 7 | 1 | 42 |
| **1DRV** | 7 | 13 | 1 | 2 | 0 | 21 | 0 | 0 | 0 | 11 | 55 |
| **1IPR** | 3 | 1 | 2 | 12 | 0 | 1 | 24 | 7 | 0 | 2 | 52 |
| **1IBE** | 1 | 0 | 0 | 11 | 1 | 1 | 9 | 18 | 1 | 0 | 42 |
| **1CCO** | 1 | 3 | 11 | 0 | 6 | 0 | 0 | 1 | 29 | 0 | 51 |
| **1VEN** | 1 | 9 | 0 | 0 | 1 | 14 | 0 | 1 | 0 | 27 | 53 |
| **Total** | 85 | 63 | 70 | 65 | 36 | 49 | 49 | 36 | 49 | 50 | 552 |

**Effectifs observés**

L'inspection à vue des données permet de se rendre compte que le nombre le plus élevé de chaque ligne se trouve sur la diagonale du tableau : cela signifie que le phénomène qui semble le plus fréquent est paradoxalement l'absence de changement. Quand il change d'entreprise, en haut de la hiérarchie, le PDG reste PDG et en bas, l'inspecteur de ventes aussi. Cependant, si on fait la somme des effectifs diagonaux, on voit que cela ne regroupe 261 cas sur un total de 552 soit 47,3% des effectifs. Plus de la moitié des changements d'entreprise correspondent donc à des changements de poste. Pour en voir la logique (qui existe, mais qui ne saute pas aux yeux sur ce tableau), il suffit de prendre la représentation géométrique de ce tableau que nous donne une analyse des correspondances (figure 1).

Sur cette figure 1, tous les intitulés des lignes (préfixés Orig pour origine) et des colonnes (préfixés Dest pour destination) sont représentés, ainsi qu'un angle droit au milieu qui marque symboliquement le centre du graphique. Chaque intitulé est représenté par un point qui, par convention, se trouve toujours à la première lettre de l'intitulé.

---

[9] Exemple adapté de *L'Expansion*, juin 1978

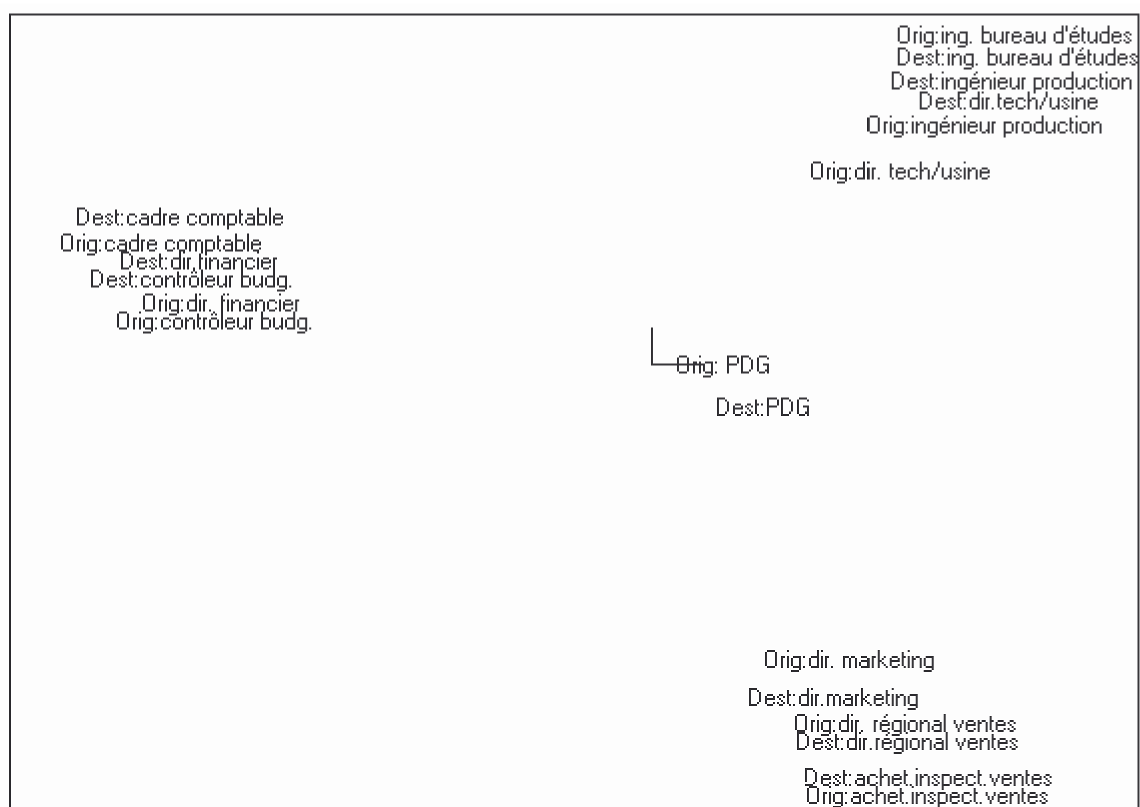



Figure 1 Représentation géométrique du tableau des carrières

Plusieurs constatations peuvent être faites :
.- les intitulés identiques (par exemple origine PDG et destination PDG) sont toujours très proches
- il y a trois groupes de points à la périphérie du graphique et un au centre qui se réduit au PDG (origine et destination)
- chacun des trois groupes est composé de postes (origines et destinations toujours proches) qui sont dans une même branche de l'entreprise : la *production* en haut à gauche (directeur technique ou d'usine, ingénieur de production ou de bureau d'études), la *fonction commerciale* en bas à droite (directeur du marketing ou directeur régional de ventes, acheteur/inspecteur de ventes), la *fonction financière* à gauche (directeur financier, contrôleur budgétaire, cadre comptable).

Ce que suggère donc la représentation géométrique du tableau des carrières c'est que les changements de postes se font à l'intérieur d'une même fonction de l'entreprise. Quand on est cadre dans le secteur de la production de l'entreprise, en cas de changement d'entreprise, on a plus de chances de se retrouver éventuellement à un autre poste mais toujours dans la production. Ce phénomène peut se repérer dans les données d'origine elles-mêmes, simplement en modifiant l'ordre des lignes et des colonnes et en regroupant les fonctions mises au jour par le graphique. Dans le tableau suivant, on a remplacé les zéros par des blancs et séparé les fonctions par des lignes pour faciliter la vision du phénomène. La direction générale a été laissée à part.

Avec une telle présentation des données d'origine, l'effet de chargement sur la diagonale est amplifié : ce sont maintenant des blocs diagonaux qui apparaissent et qui manifestent les échanges privilégiés à l'intérieur des fonctions de l'entreprise. En dehors des blocs diagonaux, les effectifs sont faibles ou nuls, sauf pour la ligne et surtout la colonne PDG qui sont spécifiques.

| Origine | | Destination PDG | DFI | CBU | CCO | DMK | DRV | VEN | DTU | IPR | IBE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Direction | 1PDG | 20 | 5 | | 1 | 3 | 1 | | 5 | 1 | |
| Finance | 1DFI | 10 | 38 | 10 | 8 | 1 | | 1 | | | |
| | 1CBU | 2 | 12 | 17 | 7 | | 2 | 1 | 1 | | |
| | 1CCO | 1 | 11 | 6 | 29 | 3 | | | | | 1 |
| Commerce | 1DMK | 22 | | | 1 | 33 | 9 | 8 | | 1 | |
| | 1DRV | 7 | 1 | | | 13 | 21 | 11 | 2 | | |
| | 1VEN | 1 | | 1 | | 9 | 14 | 27 | | | 1 |
| Production | 1DTU | 18 | 1 | 1 | 2 | | | | 34 | 14 | 9 |
| | 1IPR | 3 | 2 | | | 1 | 1 | 2 | 12 | 24 | 7 |
| | 1IBE | 1 | | 1 | 1 | | 1 | | 11 | 9 | 18 |

Effectifs observés

*Les écarts à l'indépendance : sources des attractions et des similitudes*

Pour comprendre ce que visualise la représentation géométrique fait par l'analyse factorielle, isolons le cas du PDG, origine et destination en regroupant toutes les autres lignes et toutes les autres colonnes dans une même catégorie " le reste "

| Orig. | Destination 2PDG | Le reste | Total |
|---|---|---|---|
| 1PDG | 20 | 16 | 36 |
| Le reste | 65 | 451 | 516 |
| Total | 85 | 467 | 552 |

Effectifs observés

Comme il y a 85 PDG à l'arrivée sur les 552 positions possibles, soit une proportion de 85 / 552 = 0,154 et donc un pourcentage de 15,4%. S'il y avait un échange non privilégié, indépendant de l'origine, cette règle des 15,4% s'appliquerait tout aussi bien à la ligne PDG qu'au reste. Il n'en est rien comme le manifeste le tableau des pourcentages effectués sur les lignes :

| Orig. | Destination 2PDG | Le reste | Total |
|---|---|---|---|
| 1PDG | 55,6 | 44,4 | 100 |
| Le reste | 12,6 | 87,4 | 100 |
| Total | 15,4 | 84,6 | 100 |

Pourcentages en ligne

Nous ne sommes pas dans une situation d'indépendance entre origines et destinations. S'il y avait indépendance pour le PDG, le pourcentage général de 15,4% s'appliquerait aux 36 PDG d'origine et, en multipliant cet effectif de 36 par la proportion 0,154 (ou le rapport qui la constitue 85 / 552), on aurait ce que l'on appelle un effectif théorique correspondant à l'hypothèse d'indépendance de :
36 x 85 / 552 = 5,5 individus. Or on en observe 20, ce qui fait un écart à l'indépendance de 20 – 5,5 = 14,5 individus.

Ce qui vaut pour ce cas particulier vaut en général : l'analyse des correspondances d'un tableau quelconque rapproche les lignes et les colonnes qui

sont en attraction du fait que la ligne est plus choisie par la colonne qu'en moyenne (la moyenne correspondant ici à l'effectif correspondant à l'hypothèse d'indépendance)

Prenons un autre exemple, celui de la fonction production de l'entreprise en faisant le même travail que précédemment mais en donnant comme résultat final les écarts positifs ou négatifs à la situation d'indépendance.

**Destination**

| Origine | | DTU | IPR | IBE | Le reste |
|---|---|---|---|---|---|
| Production | 1DTU | 24,7 | 7,0 | 3,8 | -35,5 |
| | 1IPR | 5,9 | 19,4 | 3,6 | -28,9 |
| | 1IBE | 6,1 | 5,3 | 15,3 | -26,6 |
| | Le reste | -36,6 | -31,6 | -22,7 | 91,0 |

**Ecarts à l'indépendance**

On voit que, en ce qui concerne les intersections de lignes et de colonnes de la fonction production, comme tous les écarts à l'indépendance sont positifs et correspondent donc à des attractions, les points lignes et colonnes sont proches dans la représentation géométrique. Mais si l'on regarde maintenant les trois lignes entre elles de la fonction production, on voit qu'elles sont semblables en terme de profil : pour toutes les colonnes, elles ont en même temps, soit des écarts positifs, soit des écarts négatifs.

Deux points de vue différents sont ainsi envisagés qui correspondent à deux formes de correspondances :
- proximités entre lignes et colonnes qui signifient une *attraction* entre les intitulés de lignes et de colonnes, repérable par un écart à l'indépendance positif ;
- proximités entre lignes entre elles (ou entre colonnes entre elles) qui signifient une *similitude* entre les intitulés de lignes (ou de colonnes), repérable par une similitude des écarts à l'indépendance (en termes de signes positifs ou négatifs)

Dans la figure 1, on repère des similitudes et des attractions entre les postes à l'intérieur de chaque fonction : similitudes entre origines (par exemple 1DMK et 1DRV), similitudes entre destinations (par exemple entre 2DFI et 2CBU), attractions entre origines et destinations à l'intérieur d'une fonction (par exemple 1PDG et 2PDG, seuls représentants de la fonction de direction générale).

*En résumé*, il faut retenir de cet exemple des carrières que l'analyse factorielle des correspondances fait la représentation géométrique d'un tableau en prenant en compte les écarts à l'indépendance du tableau (d'une manière qui sera précisée dans la suite). La notion d'indépendance dans un tableau doit donc devenir familière au lecteur[10]. Dans un tableau, l'effectif dit théorique correspondant à l'indépendance est obtenu par le produit des marges divisés par le total, c'est une manière d'appliquer le pourcentage en ligne, toutes lignes confondus, à l'effectif d'une ligne particulière. Cette hypothèse d'indépendance ne correspond à aucune théorie

[10] Cf pour un approfondissement de cette question, Philippe Cibois, Les écarts à l'indépendance. *Les écarts à l'indépendance. Techniques simples pour analyser des données d'enquêtes*, Collection "Méthodes quantitatives pour les sciences sociales", collection de livres en ligne dirigée par Alain Degenne et Michel Forsé et diffusée par la revue Sciences Humaines, 632 K. http://www.scienceshumaines.com/textesInedits/Cibois.pdf

particulière, c'est simplement l'effectif attendu quand on ne connaît que les marges du tableau qui servent d'univers de référence. L'information apportée par le tableau lui-même entraine des écarts en plus de l'indépendance (on parle alors d'attraction entre une ligne et une colonne), ou des écarts en moins (symétriquement on parle alors de répulsion ou de déficit). Dans une représentation géométrique d'un tableau, les points correspondant aux intitulés de ligne ou de colonne, s'ils sont proches manifestent une attraction. Si des lignes entre elles sont proches (respectivement des colonnes), c'est que ces lignes ont même profil d'écart à l'indépendance (positifs, négatifs ou nuls dans les mêmes colonnes), elles sont alors semblables.

## 2 Comment passer du tableau au graphique

*Décomposition du tableau de la destination des nouveaux bacheliers*

Nous allons essayer maintenant de donner une idée de la manière dont on peut passer d'un tableau à sa représentation graphique. L'exemple sera simple et, pour que la démarche soit compréhensible, c'est le principe général du passage du tableau au graphique qui sera proposé, non précisément la représentation géométrique associée à l'analyse des correspondances (dont on verra ensuite qu'elle en est cependant assez proche). L'exemple qui nous servira (destination des nouveaux bacheliers) est une simplification des données indiquant pour les bacheliers de 1996, quelle a été leur orientation dans l'enseignement supérieur l'année suivante 1996-97[11].

On a regroupé les séries du bac (en ligne) en quatre séries : Lettres (notée L), Economique et sociale (ES), Sciences (S), Technologique et pro (Tech) et les destinations en trois : Université (Univ), Classes préparatoires aux grandes écoles (CPGE) et Autres orientations à finalité professionnelle (Autres, dans lesquelles ont à mis les IUT). L'effectif des nouveaux bacheliers était cette année là de 700.000 bacheliers, population que l'on a ramené à 100 et les effectifs ont été légèrement modifiés et arrondis pour simplifier le tableau. On a le tableau suivant.

**Nouveaux bacheliers**

| | Univ | CPGE | Autr | Total |
|---|---|---|---|---|
| L | 14 | 2 | 4 | 20 |
| ES | 16 | 1 | 3 | 20 |
| S | 15 | 5 | 10 | 30 |
| Tech | 5 | 2 | 23 | 30 |
| Total | 50 | 10 | 40 | 100 |

**Effectifs ramenés à 100**

On voit que, en moyenne, la moitié des bacheliers vont à l'université, 10% dans les classes préparatoires et 40% dans les destinations à finalité professionnelles.

Puisque l'information pertinente se trouve dans les écarts à l'indépendance, c'est cette distribution marginale (50%, 10%, 40%) qui sert de référence : s'il y avait indépendance entre la série du bac et la destination, puisqu'en moyenne, la moitié des bacheliers vont à l'université, la moitié des bacheliers de la série L irait, soit 10, la moitié des ES soit 10, la moitié des S soit 15, la moitié des Tech soit 15. De la

[11] Sources : Nouveaux bacheliers dans l'enseignement supérieur 1996-97. *Repères et références statistiques 1997* du Ministère de l'Education nationale, page 171.

même façon 10% des 20 L soit 2 iraient en classes préparatoires, etc. On a le tableau correspondant à l'indépendance suivant.

Nouveaux bacheliers

| | Univ | CPGE | Autr | Total |
|---|---|---|---|---|
| L | 10 | 2 | 8 | 20 |
| ES | 10 | 2 | 8 | 20 |
| S | 15 | 3 | 12 | 30 |
| Tech | 15 | 3 | 12 | 30 |
| Total | 50 | 10 | 40 | 100 |

**Effectifs théoriques correspondant à l'indépendance**

Pour la première case (L – Université), l'effectif observé est de 14, l'effectif théorique de 10, on a donc un écart à l'indépendance positif de +4. Tous les autres écarts à l'indépendance sont calculés en faisant pour chaque case la différence *Observé* moins *Théorique*.

Nouveaux bacheliers

| | Univ | CPGE | Autr |
|---|---|---|---|
| L | 4 | 0 | -4 |
| ES | 6 | -1 | -5 |
| S | 0 | 2 | -2 |
| Tech | -10 | -1 | 11 |

**Ecarts à l'indépendance**

C'est dans ce tableau des écarts que se trouve l'information pertinente et le principe de la représentation graphique va être de tenter de donner à chaque intitulé de ligne et de colonne une valeur numérique positive ou négative unique (qui servira sur un axe du graphique). Cela semble impossible mais c'est pourtant ce que nous avons déjà pour le tableau d'indépendance. En effet chaque case du tableau est obtenue par produit des marges divisé par le total (50 x 20 / 100 pour le première case par exemple). Plutôt que de diviser le produit des marges par 100, il est possible de commencer par diviser chaque marge par 10: on a alors le tableau suivant où les marges ne sont plus des totaux mais des coefficients multiplicatifs qui permettent de calculer l'effectif correspondant à l'indépendance.

Nouveaux bacheliers

| | Univ | CPGE | Autr | Coeff. |
|---|---|---|---|---|
| L | 10 | 2 | 8 | 2 |
| ES | 10 | 2 | 8 | 2 |
| S | 15 | 3 | 12 | 3 |
| Tech | 15 | 3 | 12 | 3 |
| Coeff. | 5 | 1 | 4 | |

**Indépendance obtenue par produit de coefficients marginaux**

Dans ce tableau, chaque ligne et colonne a un coefficient spécifique. Ceci n'est possible que parce que c'est à partir des marges qu'est construit le tableau. Pour la première case, multiplier 2 par 5 redonne bien le même effectif de 10. C'est une opération analogue qu'il faudrait pouvoir faire sur le tableau des écarts à l'indépendance : trouver un jeu de coefficients qui par multiplication terme à terme (ligne par colonne), redonne les effectifs d'écarts à l'indépendance. Ceci n'est pas possible directement mais une solution qui s'en approche est possible. Soit les jeux

de coefficient suivants pour les intitulés de ligne : L=1, ES=1, S=1, Tech=-3 et pour les intitulés de colonne Univ=2, CPGE=1, Autres=-3 (nous expliquerons plus loin comment on peut trouver ces coefficients). Le tableau ci-dessous est une approximation des écarts à l'indépendance, mais ce tableau est connu par ses marges et à chaque intitulé correspond un coefficient (qui servira pour le graphique). Pour la première case par exemple 2 est le produit de 1 (coeff. L) par 2 (coeff. Univ).

Pour se rendre compte du résultat, puisqu'il s'agit d'une approximation, nous avons souligné les cases où l'approximation est la meilleure, c'est-à-dire la colonne Autres et la ligne Technique

**Nouveaux bacheliers**

| | Univ | CPGE | Autr |
|---|---|---|---|
| L | 4 | 0 | *-4* |
| ES | 6 | -1 | *-5* |
| S | 0 | 2 | -2 |
| Tech | *-10* | *-1* | *11* |

**Ecarts à l'indépendance**

| | Univ | CPGE | Autr | Coeff. |
|---|---|---|---|---|
| L | 2 | 1 | *-3* | 1 |
| ES | 2 | 1 | *-3* | 1 |
| S | 2 | 1 | *-3* | 1 |
| Tech | *-6* | *-3* | *9* | -3 |
| Coeff. | 2 | 1 | -3 | |

**Approximation**

Tous les écarts ne sont pas pris en compte, il s'en faut de la différence entre les écarts et leur approximation. Pour la première case, la différence est de 2 (4 – 2). Voici le tableau du reste :

| | Univ | CPGE | Autr | Coeff. |
|---|---|---|---|---|
| L | 2 | -1 | -1 | |
| ES | 4 | -2 | -2 | |
| S | -2 | 1 | 1 | |
| Tech | -4 | 2 | 2 | |
| Coeff. | | | | |

**Reste**

On constate que toutes les lignes et les colonnes sont proportionnelles entre elles, ce qui permet de trouver facilement des coefficients, par exemple en choisissant 1 pour L, tout le reste s'en déduit (Univ=2, CPGE=-1, Autres=-1, ES=2, S=-1, Tech=-2). Synthétisons les résultats : les écarts à l'indépendance (où se trouvent les informations pertinentes, sont la somme du tableau de l'approximation et du tableau de reste.

**Nouveaux bacheliers**

| | Univ | CPGE | Autr | | Univ | CPGE | Autr | Coef | | Univ | CPGE | Autr | Coef |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L | 4 | 0 | -4 | | 2 | 1 | -3 | 1 | | 2 | -1 | -1 | 1 |
| ES | 6 | -1 | -5 | | 2 | 1 | -3 | 1 | | 4 | -2 | -2 | 2 |
| S | 0 | 2 | -2 | | 2 | 1 | -3 | 1 | | -2 | 1 | 1 | -1 |
| Tech | -10 | -1 | 11 | | -6 | -3 | 9 | -3 | | -4 | 2 | 2 | -2 |
| | | | | Coef | 2 | 1 | -3 | | | 2 | -1 | -1 | |

**Ecarts = approximation + Reste**

Résumons les opérations faites sur les tableaux :

Tableau d'origine = Indépendance + Ecarts à l'indépendance

Ecarts = Approximation + Reste

| Tableau d'origine = Indépendance + Approximation + Reste |
|---|

On a ainsi décomposé le tableau d'origine en trois tableaux qui ont tous la propriété d'être connus par leurs marges, d'être des *faux tableaux*, c'est-à-dire que la connaissance des marges dispense de la connaissance du contenu du tableau. Toute l'analyse factorielle réside dans ce principe : on décompose un tableau d'origine en un ensemble bien ordonné de *faux tableaux* connus par leurs marges dont la somme redonne pourtant le tableau d'origine et dont les marges vont permettre une visualisation graphique.

Ensemble ordonné de tableaux : on part du tableau de départ et on en cherche une bonne approximation. En analyse des correspondances, la première approximation sous forme de tableau connu par ses marges est le tableau correspondant à l'indépendance. Cette approximation est grossière puisqu'elle laisse de côté les écarts à l'indépendance qui constituent un premier reste. On refait l'opération de recherche d'une approximation de ce premier reste (les écarts) et on peut décomposer ces écarts en leur approximation et un nouveau reste.

L'opération de décomposition est terminée et le dernier reste est déjà un faux tableau connu par ses marges : en effet, un résultat mathématique intéressant est que tout tableau est décomposable en un nombre de *faux tableaux*, que les mathématiciens appellent *tableaux de rang un,* qui dépend du nombre de lignes ou de colonne. C'est le plus petit de ces deux nombres (de lignes ou de colonnes) qui indique le nombre de tableaux au plus nécessaire pour décomposer le tableau d'origine (ce que l'on nomme le *rang* du tableau). Dans l'exemple sur la destination des bacheliers, comme il y a quatre séries en ligne et trois destinations en colonne, le rang du tableau est de trois et il peut se décomposer en trois tableaux de rang un, l'indépendance, l'approximation et le reste.

*Représentation graphique*

Pour passer à la représentation graphique, nous allons d'abord prendre le tableau d'approximation pour lequel nous disposons de coefficients pour les éléments lignes et colonnes. Nous allons disposer ces éléments sur un axe orienté dans la figure ci-dessous :

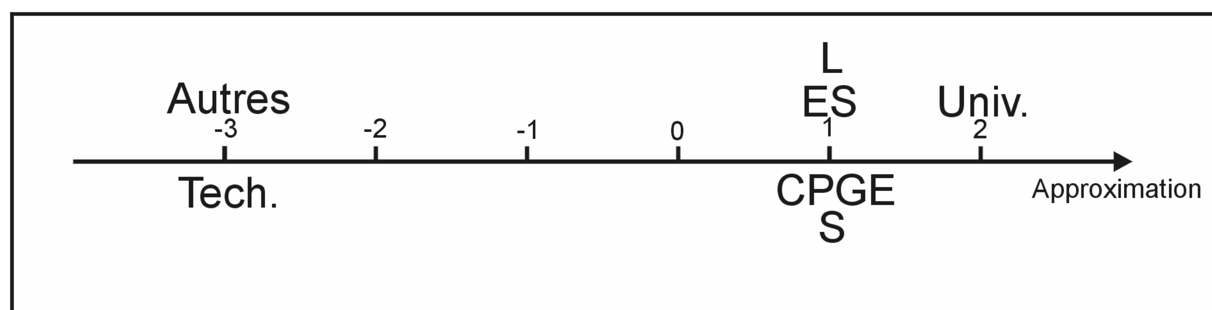



Figure 2 : Représentation graphique de l'axe du tableau approximation

Un élément seul, par exemple la ligne série technique situé à -3, ne suffit pas pour représenter une case du tableau : pour représenter la case Tech – Autres, il faut multiplier la ligne Tech par la colonne Autres, également situé à -3 et le produit, égal à 9, nous donne la valeur de *l'attraction* entre cette origine et cette destination. C'est une attraction car l'écart à l'indépendance est positif. De la même façon, nous pouvons repérer une répulsion dans le tableau en multipliant la même série technique par la destination université : le produit, -3 par 2 = -6 est négatif et indique bien une *opposition* puisque l'écart à l'indépendance est négatif.

Du côté positif de l'axe, on voit des attractions entre les séries générales (L, ES et S) et l'Université et les classes préparatoires qui, dans le tableau d'approximation constituent un bloc d'écarts tous positifs. On repère aussi les similitudes de comportement des séries générales qui sont dans le même rapport avec toutes les destinations : écarts positifs avec les destinations du côté droit (Univ et CPGE), et négatifs avec la destination du côté gauche (Autres orientations).

On peut ainsi, en utilisant les attractions, les oppositions et les similitudes procéder à une interprétation globale de l'axe qui oppose, du côté négatif, origines et destination professionnelles et, du côté positif, séries générales et l'université et classes préparatoires.

Mais le tableau d'approximation ne suffit pas, il faut aussi prendre en compte le tableau du reste pour avoir l'intégralité du tableau des écarts à l'indépendance dont ils sont la décomposition. On procède de la même façon en affectant un autre axe (disposé cette fois verticalement) et en reportant les coefficients des lignes et colonne correspondant. Le plan des deux axes est donné dans la figure 3.

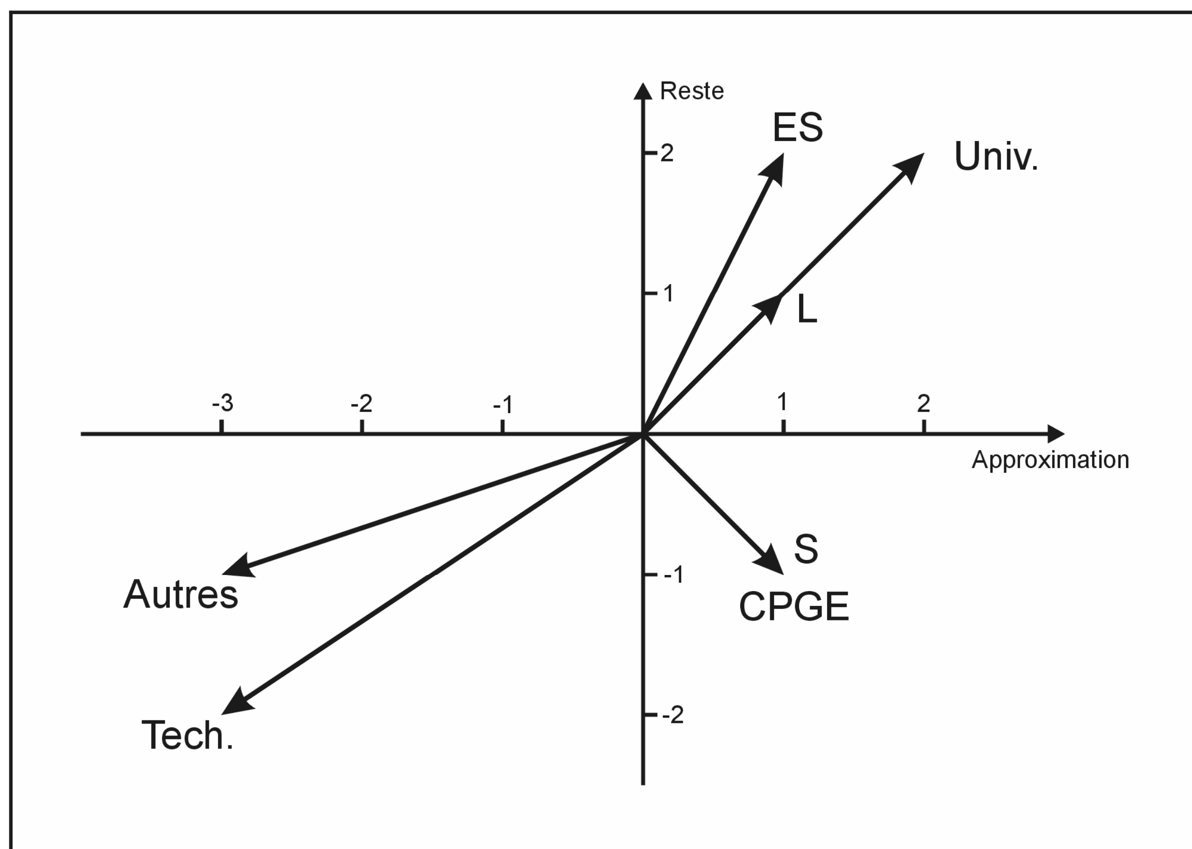



Figure 3 : Représentation graphique des écarts à l'indépendance

Les points lignes et colonnes sont représentés par des segments orientés (ou vecteurs) car les multiplications entre éléments lignes et colonnes vont maintenant devoir tenir compte de l'orientation réciproque des vecteurs. Un seul cas est simple, c'est celui de l'attraction entre la série L et l'université car les deux vecteurs ont même orientation.

Il est possible bien sûr de refaire le travail analytique précédent en multipliant les valeurs de L et d'Univ sur l'axe approximation (résultat = 2) et d'additionner ce résultat avec le même produit sur l'axe reste (même résultat, donc somme = 4) : on retrouve ainsi la valeur du tableau des écarts à l'indépendance. Cependant on peut

arriver au même résultat en travaillant dans le plan des deux axes. Il suffit de multiplier les deux vecteurs L et Univ. Comme ils ont même orientation ce produit revient à multiplier les longueurs des deux vecteurs.

Le vecteur Univ peut être considéré comme l'hypoténuse d'un triangle rectangle dont la longueur du côté est égale à 2. Il a pour longueur racine(8) soit 2 rac(2). On calcule de la même façon la longueur de L qui a pour longueur rac(2). Le produit des deux longueurs est 2 rac(2) x rac(2) = 4. On retrouve le résultat précédent.

Quand les vecteurs ne sont pas dans la même orientation, il faut tenir compte de leur angle au centre. Par exemple pour l'attraction entre ES et Univ., le produit de leurs longueur respectives (rac(5) pour ES, 2rac(2) pour Univ. soit 6,32 doit être multiplié par le cosinus de leur angle (cos(18,4349°)=0,9487) pour retrouver la valeur 6 qui est celle du tableau des écarts. Une manière plus simple de faire est de projeter orthogonalement un vecteur sur l'autre et d'utiliser la projection pour faire la multiplication : cf fig.4.

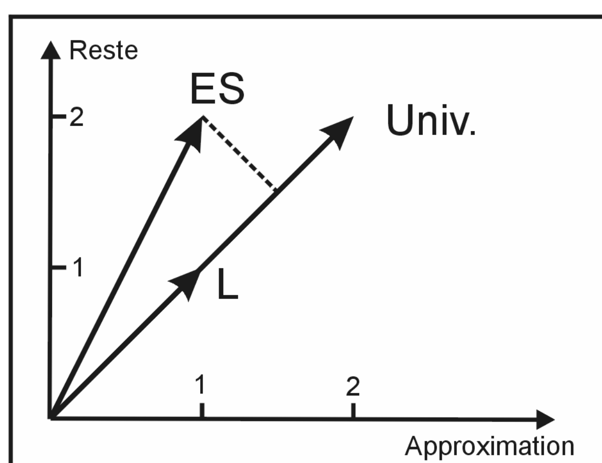



Figure 4 : projection orthogonale de ES sur Univ.

On sait que quand l'angle est faible, le cosinus est proche de 1, ce qui a peu d'influence sur le résultat. Par contre, quand l'angle s'approche de 90°, le cosinus devient proche de 0 (ou pour reprendre l'idée de la projection, celle-ci devient nulle. Par exemple, dans la figure 3, si nous projetons S sur Univ., comme l'angle est égal à 90°, la projection de S a une longueur nulle et le produit avec la longueur du vecteur Univ. est nulle également. Apparait ici un cas de figure nouveau, intermédiaire entre la conjonction, qui visualise un écart positif à l'indépendance et l'opposition qui visualise un écart négatif et qui est l'angle droit, ou quadrature, qui visualise un écart nul à l'indépendance, c'est-à-dire une situation d'indépendance.

Il y a deux cas de ce genre dans le tableau des écarts qui sont visualisés par la quadrature entre L et classes préparatoires (du fait des préparations littéraires, 10% des L vont dans ces classes, comme la moyenne) et quadrature entre S et Université (la moitié des S y vont, comme la moyenne). Ces deux cas d'écarts nuls à l'indépendance sont visualisés par les deux quadratures.

Reste le cas de figure d'opposition. Prenons par exemple l'opposition entre technique et université qui correspond à un écart négatif de -10. Pour tenir compte de l'angle, il faut projeter le vecteur Univ orthogonalement au vecteur technique, ce qui n'est possible que sur la prolongation de Tech. La projection de Univ est de sens contraire au vecteur Tech. et le résultat de la multiplication est négatif. Cf figure 5

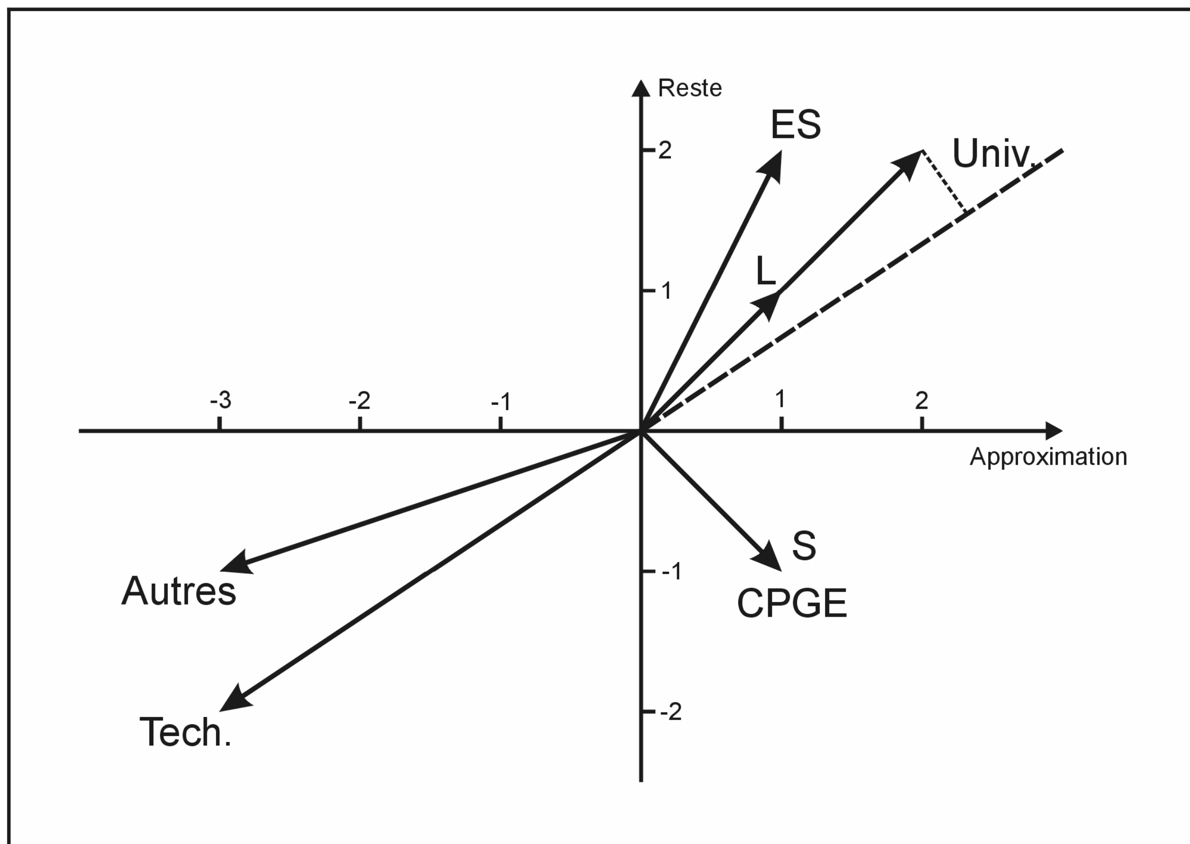


Figure 5 : projection de Univ. sur la prolongation en sens opposé de Tech.

On peut repérer ainsi trois cas de figures angulaires :
- la *conjonction* qui manifeste un écart positif à l'indépendance et qui s'interprète comme une *attraction* entre une ligne et une colonne,
- la *quadrature* qui manifeste un écart nul à l'indépendance et qui s'interprète comme une situation d'indépendance,
- l'*opposition* qui manifeste un écart négatif à l'indépendance et qui s'interprète comme une *répulsion*, un déficit entre une ligne et une colonne.

Tous les intermédiaires sont possibles entre ces cas de figure et on passe progressivement de la conjonction faible à la quasi-quadrature puis à l'opposition.

En termes de similitude, les lignes L et ES, qui sont proches, sont donc en situation voisine de conjonction, quadrature et opposition avec toutes les colonnes du tableau et seront donc en attraction avec Univ, en opposition avec Autres et, avec CPGE,soit en situation d'indépendance, soit en légère opposition.

Après avoir interprété le premier axe, on peut interpréter le plan et l'on voit trois groupes de points. Du côté gauche de l'axe, la finalité professionnelle et la série qui l'alimente (Tech. et Autres). Du côté noté comme "enseignement général", on voit que le deuxième axe (associé au tableau Reste), oppose en haut l'université et les séries qui l'alimentent et en bas la série S et les classes préparatoires. On a donc une structure en triangle : université, classes préparatoires, finalités professionnelles. L'opposition majeure, prise en compte par le premier axe oppose le professionnel au reste. Ensuite, un deuxième facteur dissocie l'université des classes préparatoires.

Le principe du passage à un tableau à sa représentation graphique est posé, il reste à voir comment il s'opère dans le cas général.

## 3 Les calculs de l'analyse factorielle

Ce que nous cherchons maintenant, c'est à partir d'un tableau quelconque, d'en trouver un jeu de coefficients pour les lignes et les colonnes qui permettent, par multiplication terme à terme, de trouver un tableau connu par ses marges. Pour montrer comment peut se faire cette recherche, nous allons utiliser un tableau à trois lignes (marquées A, B et C) et deux colonnes (I et II) : il s'agit d'un exemple choisi pour sa simplicité, mais qui ne correspond à aucune donnée précise.

| | I | II |
|---|---|---|
| A | 0 | 1 |
| B | 1 | 2 |
| C | 3 | 3 |

*Recherche de coefficients lignes et colonnes*

Examinons les colonnes du tableau : dans les deux cas, le premier élément est inférieur au deuxième, lui-même inférieur au troisième. La suite de coefficients colonnes que nous recherchons, et désormais nous appellerons ces suites de nombres des *vecteurs*, ce vecteur colonne donc, qui doit être un résumé des deux colonnes, doit avoir leur structure et doit donc ressembler à quelque chose comme (1, 2, 4) ou (1, 5, 10) mais certainement pas (10, 5, 1).

Le principe de la suite d'opérations (logiques et arithmétiques) que nous allons effectuer est ce qu'on appelle un *algorithme* : comme beaucoup d'algorithmes, il suppose une valeur de départ, même imprécise, qui sera améliorée dans la suite. Cela peut sembler inhabituel, mais ce ne l'est pas du tout, ainsi pour cet algorithme bien connu symbolisé par ce dessin :



Il s'agit du graphisme utilisé dans l'algorithme de la division : car il s'agit bien d'un algorithme puisqu'il faut un point de départ pour l'effectuer. Dans l'antique ritournelle des maitres d'antan "en dix combien de fois trois ? Il y va trois fois", que de complexité dans ce "il y va trois fois", car il s'agit bien de faire une estimation grossière du résultat final (ce qui fait d'ailleurs la difficulté de l'opération pour de jeunes enfants). Ensuite, on enchaine les opérations arithmétiques (3 x 3 = 9), les opérations logiques (9 est bien inférieur à trois, sinon il faudrait prendre une autre valeur initiale plus faible), une soustraction (10 – 9 = 1), un nouveau test (si le résultat était supérieur à 3, il faudrait prendre aussi une valeur initiale plus forte), si l'on veut une précision supplémentaire, on recommence l'opération et la règle d'arrêt ne sera pas donnée ici par un reste nul mais par une décision de l'utilisateur qui devra décider de la précision en fonction de l'utilisation en cours. Nous avons intériorisé l'algorithme, il est devenu une boite noire, mais en ouvrant cette boite, on en découvre tous les mécanismes complexes.

Nous prendrons donc comme point de départ à améliorer le vecteur colonne (1, 2, 4). Ici la suite des opérations consiste à multiplier *scalairement* le vecteur colonne à chacune des deux colonnes. Cette multiplication scalaire nous est aussi familière mais dans le registre de l'opération "facture", qui consiste, pour chacun des éléments achetés, à multiplier chacun par son prix individuel et à additionner le tout. Le résultat de la multiplication des deux vecteurs n'est pas un vecteur mais un résultat numérique sur l'échelle numérique (*scala* est l'échelle en italien).

Faisons l'opération en appelant le vecteur initial du nom de F0 et le résultat final en ligne du nom de F1:

| | I | II | F0 |
|---|---|---|---|
| A | 0 | 1 | 1 |
| B | 1 | 2 | 2 |
| C | 3 | 3 | 4 |
| F1 | 14 | 17 | |

Le premier élément de F1 s'obtient en multipliant scalairement la colonne I et F0, le détail du calcul est le suivant

| I | | F0 | | |
|---|---|---|---|---|
| 0 | x | 1 | = | 0 |
| 1 | x | 2 | = | 2 |
| 3 | x | 4 | = | 12 |
| | | | Total= | 14 |

En faisant de même pour la colonne II, on obtient le nouveau vecteur F1, constitué à partir des deux résultats. On constate que ce vecteur respecte la structure des trois lignes où le premier élément est inférieur ou à la limite égal au deuxième. Sans prétendre justifier l'algorithme, on voit qu'il intègre progressivement la structure des données du tableau. Pour continuer, il faut répéter la multiplication scalaire du vecteur F1 mais cette fois avec chacune des lignes du tableau.

| | I | II | F0 | F2 |
|---|---|---|---|---|
| A | 0 | 1 | 1 | 17 |
| B | 1 | 2 | 2 | 48 |
| C | 3 | 3 | 4 | 93 |
| F1 | 14 | 17 | | |

Pour la ligne C le détail du calcul est le suivant :

| C | | F1 | | |
|---|---|---|---|---|
| 3 | x | 14 | = | 42 |
| 3 | x | 17 | = | 51 |
| | | | Total= | 93 |

La structure de F2 est comparable à celle de F0, notre point de départ arbitraire (mais choisi avec vraisemblance), en arrondissant on peut dire que F2 a pour structure (20, 50, 90), soit, en divisant chaque élément par 20, ce qui ne modifie pas la structure (1, 2½, 4½) assez proche du point de départ. Pour pouvoir voir le phénomène avec plus de précision, examinons la structure en proportion de chacun des vecteurs. Par exemple pour F2, le premier élément 17 représente 17 / 158 = 0,108 soit 10,8%.

| | I | II | | F0 | PropF0 | F2 | PropF2 |
|---|---|---|---|---|---|---|---|
| A | 0 | 1 | | 1 | 0,143 | 17 | 0,108 |
| B | 1 | 2 | | 2 | 0,286 | 48 | 0,304 |
| C | 3 | 3 | | 4 | 0,571 | 93 | 0,589 |
| | | | Total | 7 | 1,000 | 158 | 1,000 |
| F1 | 14 | 17 | 31 | | | | |
| PropF1 | 0,452 | 0,548 | 1,000 | | | | |

On voit que de F0 à F2, la proportion du premier élément a baissé, ceux des autres a augmenté. Continuons les *itérations* de l'algorithme, c'est à dire reprenons les étapes précédentes en prenant la valeur de F2 à la place de celle de F0. Nous

multiplions scalairement chacune des colonnes du tableau par F2 et nous obtenons F3 puis à partir de F3 multiplié par chacune des lignes nous obtenons F6.

| | I | II | | F0 | PropF0 | F2 | PropF2 | F4 | PropF4 |
|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 1 | | 1 | 0,143 | 17 | 0,108 | 392 | 0,107 |
| B | 1 | 2 | | 2 | 0,286 | 48 | 0,304 | 1111 | 0,304 |
| C | 3 | 3 | | 4 | 0,571 | 93 | 0,589 | 2157 | 0,589 |
| | | | Total | 7 | 1,000 | 158 | 1,000 | 3660 | 1,000 |
| F1 | 14 | 17 | 31 | | | | | | |
| PropF1 | 0,452 | 0,548 | 1,000 | | | | | | |
| F3 | 327 | 392 | 719 | | | | | | |
| PropF3 | 0,455 | 0,545 | 1,000 | | | | | | |
| F5 | 7582 | 9085 | 16667 | | | | | | |
| PropF5 | 0,455 | 0,545 | 1,000 | Stop | | | | | |

En comparant les proportions de F2 et F4, on constate que, pour une précision de trois chiffres significatifs, les proportions sont égales sauf pour le premier élément qui passe de 10,8% à 10,7%. On voit ce qu'on appelle la *convergence* de l'algorithme qui se stabilise pour une précision donnée. Il suffit de faire une itération supplémentaire et passer de F4 à F5 pour retrouver strictement les proportions de F3. L'algorithme est terminé. Nous nous sommes affranchis de la valeur arbitraire du point de départ, les vecteurs sont maintenant *propres* aux données. Pour s'en rendre compte il suffit de changer F0 et de prendre par exemple la valeur la plus neutre possible (1, 1, 1).

| | I | II | | F0 | PropF0 | F2 | PropF2 | F4 | PropF4 | F6 | PropF6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 1 | | 1 | 0,333 | 6 | 0,115 | 128 | 0,107 | 2958 | 0,107 |
| B | 1 | 2 | | 1 | 0,333 | 16 | 0,308 | 362 | 0,304 | 8384 | 0,304 |
| C | 3 | 3 | | 1 | 0,333 | 30 | 0,577 | 702 | 0,589 | 16278 | 0,589 |
| | | | Total | 3 | 1,000 | 52 | 1,000 | 1192 | 1,000 | 27620 | 1,000 |
| F1 | 4 | 6 | 10 | | | | | | | | Stop |
| PropF1 | 0,400 | 0,600 | 1,000 | | | | | | | | |
| F3 | 106 | 128 | 234 | | | | | | | | |
| PropF3 | 0,453 | 0,547 | 1,000 | | | | | | | | |
| F5 | 2468 | 2958 | 5426 | | | | | | | | |
| PropF5 | 0,455 | 0,545 | 1,000 | | | | | | | | |

Prendre un vecteur initial quelconque a modifié tous les effectifs mais non les proportions, on voit seulement qu'il a fallu une itération supplémentaire (PropF6 = PropF4) pour arriver à la convergence de l'algorithme. De même, si on prend un point de départ (qui peut tout aussi bien être pris en ligne), complètement erroné comme (10, 5, 1), on constate que la convergence n'est pas assurée à l'itération 6. Prendre un mauvais point de départ a pour effet simplement d'augmenter le nombre d'itération. Dans une programmation en machine, on prend toujours le point de départ le plus neutre possible, soit (1, 1, 1)

Nous avons donc maintenant un couple de coefficients lignes et colonnes, des *vecteurs propres* aux données, qui expriment le mieux possible la structure du tableau, à condition qu'ils soient pris ensemble, par multiplication.

*Reconstitution du tableau d'approximation*

La reconstitution se fait donc par multiplication terme à terme des coefficients marginaux lignes et colonnes, il faut prendre les vecteurs propres (donc après convergence de l'algorithme), c'est-à-dire à l'étape 5 pour le vecteur en ligne et à l'étape 6 pour le vecteur en colonne. Se pose simplement le problème de savoir quel vecteur propre choisir, celui en effectifs ou celui en proportions ? Comme ils sont proportionnels, ils expriment tous la même structure et il en existe donc une infinité de semblables. Ici, il s'agit de faire l'approximation d'un tableau d'origine dont la somme des éléments est égale à 10 (cf. le tableau ci-dessous où les marges du tableau et son total sont calculés).

Tableau d'origine

| | I | II | Total |
|---|---|---|---|
| A | 0 | 1 | 1 |
| B | 1 | 2 | 3 |
| C | 3 | 3 | 6 |
| Total | 4 | 6 | 10 |

Approximation

| | I | II | F6 |
|---|---|---|---|
| A | 0,049 | 0,058 | 0,107 |
| B | 0,138 | 0,166 | 0,304 |
| C | 0,268 | 0,321 | 0,589 |
| F5 | 0,455 | 0,545 | 1 |

Proportion

| | I | II | F6 |
|---|---|---|---|
| A | 0,49 | 0,58 | 0,107 |
| B | 1,38 | 1,66 | 0,304 |
| C | 2,68 | 3,21 | 0,589 |
| F5 | 0,455 | 0,545 | 10 |

Multiplié par 10

On calcule d'abord l'approximation en proportion par multiplication terme à terme (par exemple pour la première case A-I 0,107 x 0,455 = 0,049), puis, pour rendre la comparaison possible, on multiplie le résultat obtenu par 10. On voit alors que cette première case est approximée par 0,49. Pour faciliter la comparaison, on à augmenté la taille des unité et l'on voit que l'approximation est plutôt "bonne". Pour la dernière ligne, pour la première colonne, il manque 3 – 2,68 = 0,32 et pour la deuxième, il y a 0, 21 en trop. Examinons toutes les erreurs en étendant le calcul par soustraction à l'ensemble : on obtient le tableau du *reste*, ce qu'il faut ajouter à l'approximation pour retrouver le tableau d'origine.

Tableau d'origine = Approximation + Reste

| | I | II | | I | II | | I | II |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 1 | A | 0,49 | 0,58 | A | -0,49 | 0,42 |
| B | 1 | 2 | B | 1,38 | 1,66 | B | -0,38 | 0,34 |
| C | 3 | 3 | C | 2,68 | 3,21 | C | 0,32 | -0,21 |

On voit sur cet exemple que l'approximation a beaucoup plus d'importance que le reste : la plus petite valeur qu'on y rencontre, 0,49 est la plus grande (en valeur absolue) du reste. L'algorithme utilisé nous a permis de décomposer un tableau en deux tableaux dont le premier est une bonne approximation du tableau d'origine.

Mais il y a plusieurs types d'algorithme, celui qui est le plus utilisé aujourd'hui est l'algorithme de l'analyse des correspondances qui, pour ne pas que les colonnes ou les lignes les plus importantes en effectif imposent le choix de l'élément prépondérant du facteur, introduit une *pondération par les marges*. A chaque pas de l'algorithme, quand un vecteur est obtenu, il est pondéré par les marges, c'est à dire divisé par elles. Reprenons l'exemple précédent en utilisant le point de départ le plus neutre possible, c'est à dire (1, 1, 1).

| | I | II | Total | F0 | F2NonPond | F2Pond |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 1 | 1 | 1 |
| B | 1 | 2 | 3 | 1 | 3 | 1 |
| C | 3 | 3 | 7 | 1 | 7 | 1 |
| Total | 4 | 6 | | | | |
| F1NonPond | 4 | 6 | | | | |
| F1Pond | 1 | 1 | | | | |

Comme on l'a vu plus haut, le résultat obtenu pour F1 est (4, 6). Il est encore ici non pondéré, le pondérer, c'est le diviser par les marges et trouver comme vecteur F1 pondéré la valeur (1, 1). Le processus se répète dans l'autre sens et en multipliant le vecteur F1 pondéré par le tableau on obtient un vecteur F2 non pondéré égal à la marge en colonne. En pondérant on retrouve le vecteur F0 de départ et l'algorithme se termine ici puisque la convergence est immédiate.

Pour la reconstitution, on se sert des vecteurs non pondérés (identiques aux marges) et le produit des marges est (à la division par le total près) identique à l'effectif théorique correspondant à l'indépendance.

Tableau d'origine

| | I | II | Total |
|---|---|---|---|
| A | 0 | 1 | 1 |
| B | 1 | 2 | 3 |
| C | 3 | 3 | 6 |
| Total | 4 | 6 | 10 |

Approximation

| | I | II | F2NP |
|---|---|---|---|
| A | 4 | 6 | 1 |
| B | 12 | 18 | 3 |
| C | 24 | 36 | 6 |
| F1NP | 4 | 6 | |

| | I | II |
|---|---|---|
| A | 0,40 | 0,60 |
| B | 1,20 | 1,80 |
| C | 2,40 | 3,60 |

Divisé par 10

Dans ce cas particulier, la première approximation correspond à l'indépendance est le reste constitue les écarts à l'indépendance.

Tableau d'origine =

| | I | II |
|---|---|---|
| A | 0 | 1 |
| B | 1 | 2 |
| C | 3 | 3 |

Indépendance +

| | I | II |
|---|---|---|
| A | 0,40 | 0,60 |
| B | 1,20 | 1,80 |
| C | 2,40 | 3,60 |

Ecart à l'indépendance

| | I | II |
|---|---|---|
| A | -0,40 | 0,40 |
| B | -0,20 | 0,20 |
| C | 0,60 | -0,60 |

Cette particularité est un des atouts de l'analyse des correspondances : la première approximation du tableau est l'indépendance ce qui veut dire que l'information pertinente se trouve dans le tableau des écarts à l'indépendance.

En résumé, nous avons vu qu'un tableau quelconque pouvait par le biais d'un algorithme être décomposé en une série de plusieurs tableaux : le premier, reconstitué par multiplication terme à terme des coefficients obtenus après convergence de l'algorithme, est une bonne approximation du tableau d'origine. Nous allons étudier maintenant les deux méthodes les plus couramment utilisées en analyse factorielle, la plus simple d'abord, l'analyse dite *en composantes principales*, puis sa variante, munie d'une pondération que nous avons déjà évoquée, l'analyse des correspondances.

## 4 L'analyse en composantes principales

*Les vecteurs propres, leur représentation graphique, les valeurs propres.*

Nous ne présentons cette méthode qu'à titre d'étape, peu pour elle-même bien qu'elle soit utilisée aussi en analyse factorielle. Pour éclairer le processus, nous allons revenir au tableau des écarts à l'indépendance de la destination des bacheliers vu plus haut et repris ci-dessous :

```
      ! Univ CPGE Autr !
------+----------------+
L     !   4    0   -4  !
ES    !   6   -1   -5  !
S     !   0    2   -2  !
Tech  ! -10   -1   11  !
------+----------------+
Ecarts à l'indépendance
```

Nous prenons comme vecteur de départ des éléments neutres, mais, pour accélérer le processus, en introduisant un signe négatif qui correspond à un colonne du tableau, comme par exemple dans la première, ce qui donne (1, 1, 1, -1) que nous multiplions une première fois avec les trois colonnes du tableau.

Pour simplifier les calculs, le tableau est mis en proportion.

Par exemple pour le premier élément de V1, il est la somme des valeurs absolues de la colonne université puisque on multiplie toujours par 1 les valeurs positives et par -1 la valeur négative.

| | Univ | Cpge | Autres | V0 |
|---|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 | 1,000 |
| ES | 0,060 | -0,010 | -0,050 | 1,000 |
| S | 0,000 | 0,020 | -0,020 | 1,000 |
| Tech | -0,100 | -0,010 | 0,110 | -1,000 |
| V1 | 0,200 | 0,020 | -0,220 | |

Avant de repartir pour multiplier les lignes par V1, nous allons lui faire subir une opération qui va neutraliser l'accroissement régulier de l'importance des vecteurs que nous avons constaté précédemment. Cette opération consiste à ramener l'importance de V1, ce que l'on appelle sa *norme*, à l'unité.

La norme d'un vecteur est ce que l'on appelle dans les cas habituels (vecteur à deux ou trois dimensions), sa longueur. Prenons pour comprendre le phénomène le cas du vecteur à deux dimensions (4, 3), donc représentable dans le plan.

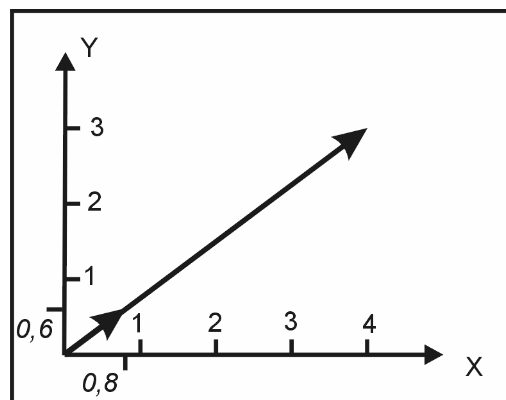



Figure 6 : vecteur (4,3) de norme 5 et vecteur (*0,8 ; 0,6*) de norme 1

Le carré de la longueur du vecteur de coordonnées (4, 3) est, du fait de théorème de Pythagore, égal à $4^2 + 3^2 = 25$. Sa longueur est donc de 5 : faire en sorte que, sans modifier la structure de ses coordonnées, qu'il soit de longueur égale à l'unité revient à diviser ses coordonnées par 5 (nouvelles coordonnées en italique). Dans le nouveau système le carré de la longueur du vecteur est :$0{,}8^2 + 0{,}6^2$ soit 0,64 + 0,36 = 1. Le carré est égal à l'unité, sa racine carrée aussi et dans le nouveau système de coordonnées, le nouveau vecteur est de norme 1 (mais est homothétique au précédent).

Le calcul se fait par les deux étapes suivantes :

| 1) Calcul de la norme de V1 | | 2) Normer V1 à l'unité | |
|---|---|---|---|
| $0{,}20^2 =$ | 0,0400 | 0,20 / *0,2980* = | 0,671 |
| $0{,}02^2 =$ | 0,0004 | 0,02 / *0,2980* = | 0,067 |
| $(-0{,}22)^2 =$ | 0,0484 | -0,22 / *0,2980* = | -0,738 |
| Somme des carrés | 0,0888 | | |
| Racine carrée | *0,2980* | | |

Nous pouvons compléter la présentation de l'algorithme par l'ajout de la ligne des carrés et de V1 normé à l'unité. Nous désignerons dans la suite les vecteurs normés à l'unité comme des vecteurs *réduits* (V1Red) et les vecteurs qui ont une norme différente de l'unité comme des vecteurs *calibrés* (V1Cal) à cette norme, ici égale à 0,298.

| | Univ | Cpge | Autres | | V0 |
|---|---|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 | | 1,000 |
| ES | 0,060 | -0,010 | -0,050 | | 1,000 |
| S | 0,000 | 0,020 | -0,020 | | 1,000 |
| Tech | -0,100 | -0,010 | 0,110 | | -1,000 |
| | | | | | |
| V1Cal | 0,200 | 0,020 | -0,220 | Somme | Norme |
| Carrés | 0,040 | 0,000 | 0,048 | 0,089 | 0,298 |
| V1Red | 0,671 | 0,067 | -0,738 | | |

NB : Trois décimales sont affichées, ce qui explique que le carré de 0,02 égal à 0,0004 soit affiché 0,000 car l'arrondi se fait au plus près.

L'importance de V1 étant maintenant réduite à l'unité, c'est ce vecteur réduit qui va maintenant servir pour continuer l'algorithme. Nous multiplions V1Red par chacune des lignes et nous obtenons V2Cal que nous réduisons également. Partant de V2Red, nous faisons une nouvelle étape qui nous donne V3 calibrés et réduits.

| | Univ | Cpge | Autres | | V0 | V2Cal | Carrés | V2Red |
|---|---|---|---|---|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 | | 1,000 | 0,056 | 0,003 | 0,318 |
| ES | 0,060 | -0,010 | -0,050 | | 1,000 | 0,077 | 0,006 | 0,431 |
| S | 0,000 | 0,020 | -0,020 | | 1,000 | 0,016 | 0,000 | 0,091 |
| Tech | -0,100 | -0,010 | 0,110 | | -1,000 | -0,149 | 0,022 | -0,840 |
| | | | | | | Somme= | 0,031 | |
| V1Cal | 0,200 | 0,020 | -0,220 | Somme | Norme | Norme= | *0,177* | |
| Carrés | 0,040 | 0,000 | 0,048 | 0,089 | *0,298* | | | |
| V1Red | 0,671 | 0,067 | -0,738 | | | | | |
| V3Cal | 0,123 | 0,006 | -0,128 | | | | | |
| Carrés | 0,015 | 0,000 | 0,016 | 0,032 | *0,178* | | | |
| V3Red | 0,690 | 0,033 | -0,723 | | | | | |

L'algorithme commence à converger : les vecteurs réduits V1 et V3 commencent à se ressembler et la norme de V3 devient très proche de la norme de V2. Cette valeur est propre au tableau et est de ce fait appelée la *valeur propre* : elle manifeste l'importance du tableau que l'on va reconstituer avec le produit terme à terme des vecteurs propres. Effectuons encore deux itérations de l'algorithme pour arriver jusqu'à V5 et voir la convergence effectuée.

| | Univ | Cpge | Autres |
|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 |
| ES | 0,060 | -0,010 | -0,050 |
| S | 0,000 | 0,020 | -0,020 |
| Tech | -0,100 | -0,010 | 0,110 |

| V0 | V2Cal | Carrés | V2Red | ***V4Cal*** | Carrés | V4Red |
|---|---|---|---|---|---|---|
| 1 | 0,056 | 0,003 | 0,318 | ***0,057*** | 0,003 | 0,318 |
| 1 | 0,077 | 0,006 | 0,431 | ***0,077*** | 0,006 | 0,435 |
| 1 | 0,016 | 0,000 | 0,091 | ***0,015*** | 0,000 | 0,085 |
| -1 | -0,149 | 0,022 | -0,840 | ***-0,149*** | 0,022 | -0,838 |
| | Somme= | 0,0315 | | | ***0,0315*** | |
| | Norme= | *0,177* | | | *0,178* | |

| | Univ | Cpge | Autres | Somme | Norme |
|---|---|---|---|---|---|
| V1Cal | 0,200 | 0,020 | -0,220 | | |
| Carrés | 0,040 | 0,000 | 0,048 | 0,0888 | *0,298* |
| V1Red | 0,671 | 0,067 | -0,738 | | |
| V3Cal | 0,123 | 0,006 | -0,128 | | |
| Carrés | 0,015 | 0,000 | 0,016 | 0,0315 | *0,178* |
| V3Red | 0,690 | 0,033 | -0,723 | | |
| ***V5Cal*** | ***0,123*** | ***0,006*** | ***-0,128*** | | |
| Carrés | 0,015 | 0,000 | 0,016 | ***0,0315*** | *0,178* |
| V5Red | 0,690 | 0,032 | -0,723 | | Stop |

V5 en effet reproduit V3 (pour la précision donnée de l'arrondi à trois décimales). Les vecteurs V5 pour les lignes et V4 pour les colonnes sont des vecteurs propres, *0,178* est la valeur propre commune à ces vecteurs propres (c'est leur norme commune). Pour reconstituer le tableau qui soit l'approximation en composantes principales du tableau de départ, il suffit de multiplier terme à terme V4 et V5. En multipliant les vecteurs réduits on a un tableau d'"importance" égale à l'unité. Pour qu'il soit calibré à la valeur propre, il suffit de multiplier chaque case par la valeur propre (on pourrait équivalemment multiplier un vecteur propre calibré, qui intègre la valeur propre, par un vecteur propre réduit).

| | Univ | Cpge | Autres | V4Red |
|---|---|---|---|---|
| L | 0,039 | 0,002 | -0,041 | 0,318 |
| ES | 0,053 | 0,002 | -0,056 | 0,435 |
| S | 0,010 | 0,000 | -0,011 | 0,085 |
| Tech | -0,103 | -0,005 | 0,108 | -0,838 |
| V5Red | 0,690 | 0,032 | -0,723 | *0,178* |

Par exemple pour la première case 0,039 = 0,318 x 0,690 x 0,178

Pour voir quelle est la valeur de cette approximation, il suffit de retrancher ce tableau au tableau de départ, ce qui nous donne le reste.

| | Tableau des écarts = | | | Approximation + | | | Reste | | |
|---|---|---|---|---|---|---|---|---|---|
| | Univ | Cpge | Autres | Univ | Cpge | Autres | Univ | Cpge | Autres |
| L | 0,040 | 0,000 | -0,040 | 0,039 | 0,002 | -0,041 | 0,001 | -0,002 | 0,001 |
| ES | 0,060 | -0,010 | -0,050 | 0,053 | 0,002 | -0,056 | 0,007 | -0,012 | 0,006 |
| S | 0,000 | 0,020 | -0,020 | 0,010 | 0,000 | -0,011 | -0,010 | 0,020 | -0,009 |
| Tech | -0,100 | -0,010 | 0,110 | -0,103 | -0,005 | 0,108 | 0,003 | -0,005 | 0,002 |

Par exemple pour la première case 0,040 = 0,039 + 0,001

Ne serait-ce qu'en regardant les valeurs absolues, on voit que le reste a peu d'importance et que l'approximation exprime la plus grande part de l'information

contenue dans le tableau des écarts. Ceci vient de ce que le reste est le dernier tableau de la décomposition qui n'en comporte que deux pour les écarts. En effet, le tableau d'origine à trois lignes et quatre colonnes est de rang 3, il se décompose au maximum en trois tableaux connus par leurs marges. Le tableau des écarts à l'indépendance (Observés – théoriques), est déjà le résultat d'une soustraction par un tableau de rang 1 (effectifs théoriques, tableau connu par ses marges) : il n'est donc plus que de rang 2 et se décompose en deux tableaux de rang 1. Quand on a le premier (l'approximation), le 2$^{e}$ (le reste) est déjà de rang 1.

Pour trouver les vecteurs propres et la valeur propre de ce reste, il suffit de recommencer l'algorithme en prenant comme vecteur initial (1, 1, -1, 1) en respectant les signes de la première colonne du reste pour accélérer la convergence qui est d'ailleurs rapide.

Entre les vecteurs V2 et V4, il n'y a pas de différences et nous pouvons nous arrêter à ce niveau. Si nous faisons la reconstitution du reste à partir de ses vecteurs V3 et V4 réduits et de la valeur propre 0,029. Nous constatons que c'est exactement le reste lui-même que nous reconstituons, ce qui montre bien que la décomposition factorielle est terminée.

| | Univ | Cpge | Autres | | V0 | V2Cal | Carrés | V2Red | V4Cal | Carrés | V4Red | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L | 0,001 | -0,002 | 0,001 | | 1 | 0,002 | 0,0000 | 0,077 | 0,002 | 0,0000 | 0,077 | |
| ES | 0,007 | -0,012 | 0,006 | | 1 | 0,015 | 0,0002 | 0,525 | 0,015 | 0,0002 | 0,525 | |
| S | -0,010 | 0,020 | -0,009 | | -1 | -0,024 | 0,0006 | -0,819 | -0,024 | 0,0006 | -0,819 | |
| Tech | 0,003 | -0,005 | 0,002 | | 1 | 0,006 | 0,0000 | 0,218 | 0,006 | 0,0000 | 0,218 | |
| | | | | | | Somme= | 0,0009 | | | 0,0009 | | |
| V1Cal | 0,021 | -0,039 | 0,018 | Somme | Norme | Norme= | *0,029* | | | *0,029* | Stop | |
| Carrés | 0,000 | 0,002 | 0,000 | 0,0023 | *0,048* | | | | | | | |
| V1Red | 0,437 | -0,816 | 0,379 | | | | | | | | | |
| V3Cal | 0,013 | -0,024 | 0,011 | | | | | | | | | |
| Carrés | 0,000 | 0,001 | 0,000 | 0,0009 | *0,029* | | | | | | | |
| V3Red | 0,437 | -0,816 | 0,379 | | | | | | | | | |

Algorithme de recherche des vecteurs propres et valeur propre pour le reste

| | Univ | Cpge | Autres | V4Red |
|---|---|---|---|---|
| L | 0,001 | -0,002 | 0,001 | 0,077 |
| ES | 0,007 | -0,012 | 0,006 | 0,525 |
| S | -0,010 | 0,020 | -0,009 | -0,819 |
| Tech | 0,003 | -0,005 | 0,002 | 0,218 |
| V3Red | 0,437 | -0,816 | 0,379 | *0,029* |

Reconstitution du reste par multiplication

Nous pouvons maintenant procéder à la représentation géométrique des données en utilisant les vecteurs propres après convergence de l'algorithme pour le premier facteur, qui sera mis sur l'axe horizontal, et pour le deuxième, sur l'axe vertical. Rappelons que pour représenter un tableau, le coefficient ligne ou colonne ne suffit pas et que c'est l'ensemble des conjonctions, quadrature et oppositions entre lignes et colonnes qui le représente.

Deux types de vecteurs sont utilisables, les vecteurs calibrés (figure 7) ou les vecteurs réduits (figure 8). On constate que dans la figure 7, le deuxième axe a peu d'extension alors qu'avec les vecteurs réduits, les deux axes ont la même extension.

La figure 7 où l'approximation (axe horizontal) apporte beaucoup plus d'information que le reste (axe vertical) représente bien cette différence tandis que la figure 8, en donnant la même valeur aux deux axes la masque. C'est la raison qui fait que dans la suite on utilisera toujours les vecteurs calibrés : ce sont eux qui sont donnés dans les logiciels et utilisés pour les graphiques.

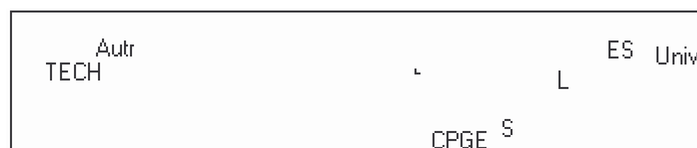



Figure 7 : plan des vecteurs calibrés

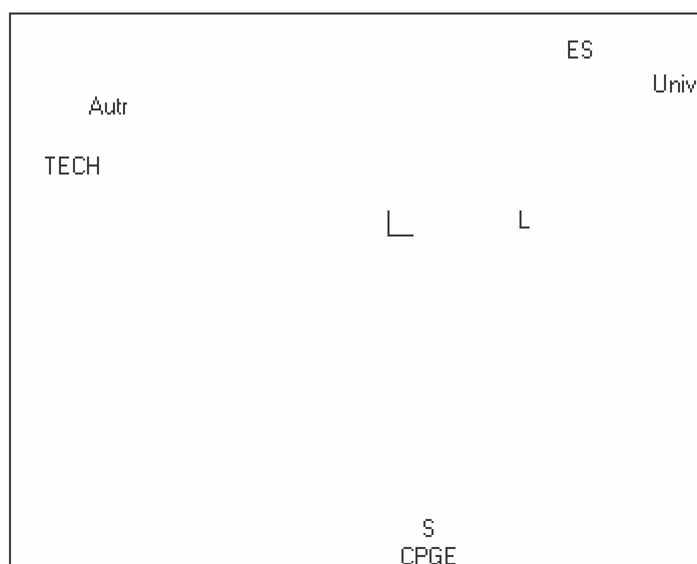



Figure 8 : plan des vecteurs réduits

Cet aspect de l'importance différente entre l'axe horizontal qui correspond au premier tableau (et on l'appelle de ce fait premier axe), nous introduit au concept clé de *contribution* qui permet de quantifier cette importance.

*Les contributions*

L'approximation et le reste (ou pour parler dans le cas général, le premier facteur et le deuxième), n'ont pas la même importance. Pour quantifier cette importance, on étend la notion de norme utilisée pour les facteurs aux tableaux : le tableau de départ et les tableaux correspondant aux deux facteurs. Comme pour un vecteur, l'importance d'un tableau sera quantifiée par la somme des carrés de ses éléments. En faisant l'exercice pour l'ensemble des tableaux on arrive au résultat suivant où l'on a mis dans les marges les totaux en ligne et en colonne (et pour le total) des sommes des carrés.

| | Marges tableau des écarts = | | | | Marges approximation + | | | | Marges reste | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Univ | Cpge | Autres | Total | Univ | Cpge | Autres | Total | Univ | Cpge | Autres | Total |
| L | 0,00160 | 0,00000 | 0,00160 | 0,0032 | 0,00152 | 0,00000 | 0,00167 | 0,0032 | 0,00000 | 0,00000 | 0,00000 | 0,0000 |
| ES | 0,00360 | 0,00010 | 0,00250 | 0,0062 | 0,00284 | 0,00001 | 0,00311 | 0,0060 | 0,00004 | 0,00016 | 0,00003 | 0,0002 |
| S | 0,00000 | 0,00040 | 0,00040 | 0,0008 | 0,00011 | 0,00000 | 0,00012 | 0,0002 | 0,00011 | 0,00038 | 0,00008 | 0,0006 |
| Tech | 0,01000 | 0,00010 | 0,01210 | 0,0222 | 0,01056 | 0,00002 | 0,01158 | 0,0222 | 0,00001 | 0,00003 | 0,00001 | 0,0000 |
| Total | 0,0152 | 0,0006 | 0,0166 | *0,0324* | 0,0150 | 0,0000 | 0,0165 | *0,0315* | 0,0002 | 0,0006 | 0,0001 | *0,0009* |

Par exemple pour la première case :
- dans le tableau des écarts, la valeur $0,04^2 = 0,160000$
- dans l'approximation, la valeur $0,039^2 = 0,001521$
- dans le reste, la valeur $0,001^2 = 0,000001$.

Sur cet exemple on vérifie qu'au niveau des cases du tableau, il n'y a pas égalité des sommes de carrés (alors que c'était le cas au niveau des proportions) mais que la décomposition additive se fait au niveau des marges et du total. Pour la première colonne (Université) par exemple, le total 0,0152 du tableau des écarts se décompose en 0,0150 dans l'approximation et 0,0002 dans le reste.

Enfin, au niveaux des totaux, on constate que le total de 0,0324 est presque entièrement pris en compte par l'approximation avec 0,0315 et que le reste n'en exprime que 0,0009. Cette décomposition additive 0,0324 = 0,0315 + 0,0009 permet de chiffrer par un pourcentage l'importance réciproque de la contribution de l'approximation et du reste à la décomposition de la somme des carrés. Si 100% correspond à 0,0324, l'approximation apporte 0,0315 / 0,024 = 0,972 en proportion soit 97,2% en pourcentage, le reste n'en apporte que le complément soit 2,8%. Ces valeurs de 0,0315 et 0,0009 qui expriment l'importance du tableau lui sont propres et sont appelées aussi *valeurs propres*. Dans l'algorithme d'extraction des vecteurs propres, elles apparaissent aussi comme étant la somme des carrés de chacun des vecteurs propres. Le pourcentage que chaque valeur propre représente par rapport au total est appelé le *taux d'explication* (ou pourcentage d'explication, ou taux d'inertie, ou contribution) du facteur. Comme les graphiques sont faits avec des vecteurs propres calibrés à la valeur propre, on peut juger visuellement de l'importance réciproque des deux axes.

Précision de vocabulaire : de même qu'il y a deux types de vecteurs propres : calibrés et réduits, de même les valeurs propres se présentent sous deux formes, la somme des carrés (c'est la valeur propre qui est affichée par les programmes) et sa racine carrée ou norme du tableau ou du vecteur). La première (somme des carrés) est souvent notée par la lettre grecque lambda ($\lambda$), la deuxième (norme) par ksi ($\xi$).

La décomposition additive des totaux de lignes et de colonnes va permettre aussi de chiffrer la contribution d'une ligne (ou d'une colonne) à l'ensemble et de deux façons différentes. Prenons l'exemple des classes préparatoires dont la somme des carrés globale est de 0,0006 qui est pris en compte en presque totalité dans le deuxième facteur. Il ne s'agit pas d'une prise en compte à 100% comme le laisse croire l'arrondi à la $4^e$ décimale. Si l'on prend davantage de décimales, on a les résultats suivants où l'on prend deux points de vue, d'abord celui de l'ensemble des facteurs :

| CPGE | Som.Car. | Prop. |
|---|---|---|
| F1 | 0,0000329 | 0,055 |
| F2 | *0,0005671* | *0,945* |
| Total | 0,0006000 | 1 |

Alors qu'en moyenne, le deuxième facteur ne représente que 3% de la somme des carrés, pour les classes préparatoires, le $2^e$ facteur représente 94,5% de la contribution, ce qui veut dire que le premier facteur a ignoré les classes préparatoires. Cette information se retrouve si l'on regarde l'ensemble des contributions au deuxième facteur cette fois :

| F2 | | |
|---|---|---|
| Univ | 0,0001625 | 0,191 |
| Cpge | *0,0005671* | *0,666* |
| Autres | 0,0001225 | 0,144 |
| Total | 0,0008520 | 1 |

La même somme de carrés (mise en italique), rapportée aux autres sommes de carrés des autres colonnes montre que les classes préparatoires contribuent à 66,6% du total (c'est-à-dire de la valeur propre du 2$^e$ facteur).

Dans les logiciels, ces informations sont données de la manière suivante :
1) ce sont les vecteurs propres calibrés à la valeur propre (qui servent pour le graphique) qui sont donnés : pour gagner de la place, ils sont donnés multipliés par 1000 (par exemple Université dans F1 : 0,123 est affiché 123) ;
2) les sommes des carrés ne sont pas données, sauf pour la valeur propre (mais peuvent être calculées facilement en élevant l'élément du vecteur propre au carré) mais la proportion par rapport au total (soit du facteur, soit du tableau d'origine) est donnée non en pourcentage mais en pour mille (c'est donc encore la proportion multipliée par mille). Etudions en détail une sortie de logiciel (Trideux) : j'ai laissé le texte du logiciel en caractères d'origine.

```
ACP : Analyse en composantes principales des écarts
**************************************************
La somme des carrés est de : 0.032400
Facteur   1 Valeur propre =  0.031548 Pourcentage du total = 97.4
Facteur   2 Valeur propre =  0.000852 Pourcentage du total =  2.6
```

[on vérifie que la somme des valeurs propres est égale à la somme des carrés totale]

```
Somme des Cos2 pour les facteurs affichés (QLT)
```

[on utilise le terme de cos² pour désigner la proportion de somme de carrés répartie sur chaque facteur, par exemple pour les classes préparatoires 55 pour mille pour F1 et 945 pour F2. Si l'on parle de cosinus, c'est que c'est une interprétation possible en termes d'angles. Quand on n'a que deux facteurs, la somme de ces contributions vaut 1000 : quand on a plus de deux facteurs, et que l'on néglige les derniers, on donne souvent la somme pour les premiers facteurs qui sont affichés, et cette proportion est considérée comme un indice de la *qualité de la représentation* (en abrégé QLT) de la ligne ou de la colonne. Ici, si on ne prenait qu'un facteur la qualité de la représentation des classes préparatoires, égale à 55 serait mauvaise.]

```
Coordonnees factorielles (F= )
Contributions pour la variable(COS2) et contributions pour le facteur(CPF)
Lignes du tableau
*---*----*------*----*----*------*----*----*
ACT.  QLT!   F=1 COS2  CPF!   F=2 COS2  CPF!
*---*----*------*----*----*------*----*----*
L    1000!    57  998  101!     2    2    6!
ES   1000!    77  962  189!    15   38  275!
S    1000!    15  285    7!   -24  715  671!
TECH 1000!  -149  998  702!     6    2   48!
*---*----*------*----*----*------*----*----*
Moy. 1000!        974  250!         26  250!
*---*----*------*----*----*------*----*----*
```

Modalites en colonne

| ACT. | QLT! | F=1 | COS2 | CPF! | F=2 | COS2 | CPF! |
|---|---|---|---|---|---|---|---|
| Univ | 1000! | 123 | 989 | 477! | 13 | 11 | 190! |
| CPGE | 1000! | 6 | 55 | 1! | -24 | 945 | 666! |
| Autr | 1000! | -128 | 993 | 522! | 11 | 7 | 144! |
| Moy. | 1000! | | 974 | 333! | | 26 | 333! |

Sous la rubrique F= se trouvent les vecteurs propres des lignes ou des colonnes pour les facteurs 1 et 2. Ce sont les seuls éléments des résultats qui peuvent être de signe négatif.

CPF signifie *Contribution par facteur* : c'est l'indicateur qui est le plus utilisé pour l'interprétation, c'est la proportion, pour un facteur, de la somme de carrés apporté par chaque ligne ou colonne. La somme des CPF vaut mille (à l'arrondi près). Les COS2 sont sommés en ligne, les CPF sommés en colonne.

Pour un facteur donné, la moyenne des COS2 est le pourcentage d'explication du facteur dans son ensemble, c'est-à-dire le pourcentage de la valeur propre (exprimé aussi en millièmes). Pour les CPF, la valeur moyenne ne dépend que du nombre de lignes ou de colonnes, par exemple, pour les lignes du tableau, si chaque ligne apportait la même contribution, comme il n'y a que quatre lignes, chacune apporterait 1000 / 4 doit 250 pour mille. Ces valeurs moyennes permettent de voir l'apport spécifique d'un élément. Par exemple pour les classes préparatoires au 2$^{e}$ facteur, la contribution par rapport au total de 945, beaucoup plus grande que la moyenne de 26 montre bien que ce facteur a bien pris en compte les Cpge. De la même manière, la contribution par rapport au facteur CPF = 666, supérieure à la moyenne de 333 (puisqu'il y a trois colonnes) montre bien l'importance des classes préparatoires dans ce facteur.

Comme les informations apportés par les COS2 et les CPF sont souvent redondantes, on peut souvent prendre la version simplifiée et ne considérer que les CPF : par exemple le résultat par défaut dans Trideux est :

Coordonnees factorielles (F= ) et contributions pour le facteur (CPF)

Lignes du tableau

| ACT. | F=1 | CPF | F=2 | CPF |
|---|---|---|---|---|
| L | 57 | 101 | 2 | 6 |
| ES | 77 | 189 | 15 | 275 |
| S | 15 | 7 | -24 | 671 |
| TECH | -149 | 702 | 6 | 48 |
| * | * | *1000* | | *1000* |

Modalites en colonne

| ACT. | F=1 | CPF | F=2 | CPF |
|---|---|---|---|---|
| Univ | 123 | 477 | 13 | 190 |
| CPGE | 6 | 1 | -24 | 666 |
| Autr | -128 | 522 | 11 | 144 |
| * | * | *1000* | | *1000* |

On ne trouve ici que les vecteurs propres et leur contribution par facteur : cette présentation compactée permet une interprétation suffisante des données. Par exemple pour le premier facteur et en considérant les lignes du tableau, on peut négliger la faible contribution de S (8 pour mille) et parler d'une opposition entre technique (coordonnée négative) et Lettres ainsi qu'ES (côté positif). De la même façon pour les colonnes, l'opposition se fera entre universités et autres orientations, les classes préparatoires ayant une contribution négligeable. Pour le 2^e^ facteur, on l'a déjà noté, ce sont les classes préparatoires (et S pour les lignes) qui contribuent le plus à ce facteur en s'opposant au reste.

Avec les vecteurs propres et leur représentation graphique, les valeurs propres, les contributions, nous avons vu l'essentiel des concepts de l'analyse factorielle mais il reste à étudier une technique dérivée de l'algorithme d'obtention des vecteurs propres qui est d'une grande utilité, la technique des *éléments supplémentaires*.

*Les éléments supplémentaires*

Le principe des éléments supplémentaires est que quand la convergence de l'algorithme d'extraction des vecteurs propres est terminée, une itération supplémentaire redonne les mêmes valeurs. La multiplication du vecteur réduit avec les mêmes lignes (ou les mêmes colonnes) redonnera les mêmes vecteurs calibrés. On introduit à ce moment une ligne (ou une colonne) supplémentaire. Si cette ligne ou colonne était strictement identique à une ligne ou colonne existante, le résultat de la multiplication serait identique. Si elle est légèrement différente, le résultat de la multiplication sera légèrement différent. Si nous prenons une ligne ou colonne quelconque, son résultat sera proche de la ligne ou colonne qui lui ressemble le plus.

Reprenons l'exemple du bac au niveau du premier facteur où nous utilisons comme tableau de départ les écarts à l'indépendance. Créons une ligne supplémentaire qui additionne les écarts des séries L et ES

| | Univ | Cpge | Autres |
|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 |
| ES | 0,060 | -0,010 | -0,050 |
| S | 0,000 | 0,020 | -0,020 |
| Tech | -0,100 | -0,010 | 0,110 |
| | | | |
| L+ES | 0,100 | -0,010 | -0,090 |

Nous prenons comme vecteur initial, non un vecteur quelconque, mais le dernier vecteur réduit (avec toutes ses décimales dans le calcul, même s'il n'est affiché qu'avec trois décimales). En le multipliant avec chacune des lignes, on retrouve un vecteur V1 Calibré identique à ce que nous obtenions déjà. Cependant nous ajoutons une ligne supplémentaire, la ligne L+ES. Sa multiplication avec le vecteur initial V0Red donne comme résultat 0,134 comme élément du vecteur propre (et donc comme position en x sur le graphique). Nous pouvons aussi calculer une contribution, fictive évidemment, qui serait celle d'une colonne qui lui ressemblerait en ayant la même coordonnée factorielle. On calcule donc son carré et ce qu'elle représente (en pour mille) par rapport au total soit près de la moitié (56,7%)

| | Univ | Cpge | Autres | V1Cal | Carrés | CPF |
|---|---|---|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 | 0,057 | 0,003 | 101 |
| ES | 0,060 | -0,010 | -0,050 | 0,077 | 0,006 | 189 |
| S | 0,000 | 0,020 | -0,020 | 0,015 | 0,000 | 7 |
| Tech | -0,100 | -0,010 | 0,110 | -0,149 | 0,022 | 702 |
| | | | | *Somme=* | *0,0315* | **1000** |
| L+ES | 0,100 | -0,010 | -0,090 | 0,134 | 0,018 | 567 |
| | | | | | | |
| V0Red | 0,690 | 0,032 | -0,723 | | | |

Comme les éléments supplémentaires sont calculés une fois la détermination des vecteurs propres faite, leur nombre est indifférent pour l'analyse qui n'est pas modifiée par leur présence ou leur absence. Ici on a mis une ligne supplémentaire, on pourrait mettre aussi une ou plusieurs colonnes. Le but est de pouvoir mettre en supplémentaire des lignes ou colonnes qui ont des rapports avec les données (comme le regroupement fait ici qui rassemble le secteur lettres et sciences humaines) mais qui sont cependant non incluses dans le tableau d'origine. Nous donnerons dans la suite de nombreux exemples d'utilisation.

Nous en avons terminé avec le détail de l'analyse en composante principale, il reste à voir techniquement quelles sont les modifications apportées par l'analyse des correspondances

## 5 L'analyse des correspondances

L'idée de base de la modification apportée par l'analyse des correspondances est que si on juge dans un tableau qu'une ligne ou une colonne doit être présente, même si elle ne concerne que peu d'individus, il faut faire en sorte que l'information présente soit prise en compte, mise sur le même pied que l'information des lignes ou colonnes à plus fort effectif.

Reprenons les données d'origine de la destination des bacheliers :

| | Univ | Cpge | Autres | Total |
|---|---|---|---|---|
| L | 14 | 2 | 4 | 20 |
| ES | 16 | 1 | 3 | 20 |
| S | 15 | 5 | 10 | 30 |
| Tech | 5 | 2 | 23 | 30 |
| Total | 50 | 10 | 40 | 100 |

La colonne des Classes préparatoires, a un faible effectif qui ne représente que 10% du total, mais c'est précisément parce qu'elle à un faible effectif, par le jeu de la sélection, qu'elle a une grande importance sociale. Pour que cette spécificité soit bien respectée dans l'algorithme d'extraction des vecteurs propres, une nouvelle étape va être ajoutée entre le vecteur calibré et le vecteur réduit, il s'agit d'une pondération par les marges.

Sur le même tableau (en proportion, pour des raisons de commodité) commençons l'algorithme par un vecteur V0 composé de 1. En le multipliant à chacune des colonnes, cela revient à faire la somme des éléments et on retrouve la même valeur que le total (V1 Calibré Non pondéré). Le pondérer, c'est le diviser par

le total de la colonne, identique et l'on revient, après pondération (V1 Calibré Pondéré) à 1.

| | Univ | Cpge | Autres | Total | V0 |
|---|---|---|---|---|---|
| L | 0,14 | 0,02 | 0,04 | 0,20 | 1 |
| ES | 0,16 | 0,01 | 0,03 | 0,20 | 1 |
| S | 0,15 | 0,05 | 0,10 | 0,30 | 1 |
| Tech | 0,05 | 0,02 | 0,23 | 0,30 | 1 |
| Total | 0,50 | 0,10 | 0,40 | 1 | |
| V1CalNpond | 0,50 | 0,10 | 0,40 | | |
| V1CalPond | 1 | 1 | 1 | | |

Si l'on repart dans l'autre sens en se servant du vecteur calibré pondéré, on arrivera aussi à une colonne de 1 identique à V0 dans V2

| | Univ | Cpge | Autres | Total | V0 | V2CalNpond | V2CalPond |
|---|---|---|---|---|---|---|---|
| L | 0,14 | 0,02 | 0,04 | 0,20 | 1 | 0,20 | 1 |
| ES | 0,16 | 0,01 | 0,03 | 0,20 | 1 | 0,20 | 1 |
| S | 0,15 | 0,05 | 0,10 | 0,30 | 1 | 0,30 | 1 |
| Tech | 0,05 | 0,02 | 0,23 | 0,30 | 1 | 0,30 | 1 |
| Total | 0,50 | 0,10 | 0,40 | 1 | | | |
| V1CalNpond | 0,50 | 0,10 | 0,40 | | | | |
| V1CalPond | 1 | 1 | 1 | | | | |

Pour reconstituer la première approximation du tableau d'origine, on se sert du produit terme à terme des vecteurs non pondérés V1 et V2, ce qui revient à faire le produit des marges et ce qui conduit au résultat fondamental suivant : en analyse des correspondances, la première approximation n'est autre que le tableau correspondant à l'hypothèse d'indépendance. Ce résultat est fondamental car il explique en grande partie l'efficacité de l'analyse des correspondances car les facteurs suivants vont décomposer ce qui reste, c'est-à-dire les écarts à l'indépendance, l'information pertinente d'un tableau. Pour cette raison, ce facteur initial est passé sous silence, numéroté zéro et le premier facteur en analyse des correspondances est le résultat de la recherche des vecteurs propres sur les écarts à l'indépendance.

*Recherche du premier facteur.*

Pour le premier facteur on part donc des écarts à l'indépendance (en proportion) et l'on prend comme vecteur initial V0 un facteur neutre qui, pour accélérer la convergence respecte les signes d'une colonne, par exemple la première. On le multiplie scalairement avec les différentes colonnes et on obtient le vecteur V1 calibré, comme en analyse en composantes principales mais avant pondération (noté V1CNP c'est-à-dire calibré non pondéré). La pondération consiste à diviser chaque élément du vecteur par la marge d'origine et l'on obtient le vecteur calibré pondéré (V1CPnd).

La norme du vecteur est calculée en tenant compte de la pondération. Les éléments du vecteur calibré sont élevés au carré et divisé par la pondération (ce qui revient à multiplier entre eux les éléments pondérés et non pondérés). De la somme des carrés (colonne CarPnd, carrés pondérés), valeur propre notée habituellement lambda, on calcule la racine carrée ce qui nous donne la norme.

La norme obtenue nous permet de réduire (normer) les vecteurs calibrés. En divisant chaque élément des vecteurs non pondérés ou pondérés, on obtient leur

équivalent réduit : V1RNP (Réduit non pondéré) et V1RPnd (Réduit pondéré). C'est ce dernier vecteur qui sert de point de départ pour une nouvelle étape.

Le processus est résumé sous forme d'équivalent d'un tableur. Les chiffres sont donnés avec une précision limitée mais les calculs sont faits avec toute la précision possible. Pour éviter de recommencer trop de calculs (et pour avoir une disposition qui tienne sur une page), chaque itération est séparée et le vecteur initial de la deuxième itération (mis à l'emplacement de V0 dans la première) a été recopié (en valeur dans le tableur) à partir du résultat le plus à droite de la première itération (V2RedPond). Il en est de même pour la 3$^{e}$ itération où la convergence pour la précision de 2 chiffres est obtenue car le Vecteur V6 (calibré pondéré) est identique au vecteur V4. Ce sont les vecteurs calibrés et pondérés qui sont donnés dans les logiciels usuels.

*Reconstitution de l'approximation*

Comme en composantes principales, on peut multiplier des vecteurs réduits (mais il faut multiplier le résultat par la norme) ou des vecteurs calibrés (mais il faut alors diviser). Si on utilise des vecteurs non calibrés, il n'y a rien à faire de plus, mais si on utilise des vecteurs calibrés, il faut multiplier par la pondération correspondante. Comme ce sont les vecteurs calibrés et pondérés qui sont fournis par les programmes, il faut donc, pour chaque multiplication terme à terme des éléments des vecteurs propres, multiplier par les pondérations correspondantes et diviser par la norme (qui est la racine carrée de la valeur propre lambda donnée par les programmes)

Itération 1

| | Univ | Cpge | Autres | Pond | V0 | V2CNP | V2CPnd | CarPnd | V2RNP | V2RPnd |
|---|---|---|---|---|---|---|---|---|---|---|
| L | 0,04 | 0,00 | -0,04 | 0,20 | 1,00 | 0,08 | 0,42 | 0,04 | 0,16 | 0,81 |
| ES | 0,06 | -0,01 | -0,05 | 0,20 | 1,00 | 0,11 | 0,55 | 0,06 | 0,21 | 1,05 |
| S | 0,00 | 0,02 | -0,02 | 0,30 | 1,00 | 0,03 | 0,11 | 0,00 | 0,06 | 0,21 |
| Tech | -0,10 | -0,01 | 0,11 | 0,30 | -1,00 | -0,23 | -0,75 | 0,17 | -0,44 | -1,45 |
| Pond | 0,50 | 0,10 | 0,40 | | | | Somme | 0,2695 | | |
| V1CNP | 0,20 | 0,02 | -0,22 | | | | Norme | 0,5191 | | |
| V1CPnd | 0,40 | 0,20 | -0,55 | Somme | Norme | | | | | |
| CarPnd | 0,08 | 0,00 | 0,12 | 0,2050 | 0,4528 | | | | | |
| V1RNP | 0,44 | 0,04 | -0,49 | | | | | | | |
| V1RPnd | 0,88 | 0,44 | -1,21 | | | | | | | |

Itération 2

| | Univ | Cpge | Autres | Pond | V2RPnd | V4CNP | **V4CPnd** | CarPnd | V4RNP | V4RPnd |
|---|---|---|---|---|---|---|---|---|---|---|
| L | 0,04 | 0,00 | -0,04 | 0,20 | 0,81 | 0,08 | **0,42** | 0,04 | 0,16 | 0,81 |
| ES | 0,06 | -0,01 | -0,05 | 0,20 | 1,05 | 0,11 | **0,57** | 0,06 | 0,22 | 1,09 |
| S | 0,00 | 0,02 | -0,02 | 0,30 | 0,21 | 0,03 | **0,09** | 0,00 | 0,05 | 0,17 |
| Tech | -0,10 | -0,01 | 0,11 | 0,30 | -1,45 | -0,23 | **-0,75** | 0,17 | -0,43 | -1,44 |
| Pond | 0,50 | 0,10 | 0,40 | | | | Somme | 0,2721 | | |
| V3CNP | 0,24 | 0,01 | -0,25 | | | | Norme | 0,5217 | | |
| V3CPnd | 0,48 | 0,08 | -0,62 | Somme | Norme | | | | | |
| CarPnd | 0,12 | 0,00 | 0,16 | 0,2719 | 0,5215 | | | | | |
| V3RNP | 0,46 | 0,02 | -0,48 | | | | | | | |
| V3RPnd | 0,92 | 0,16 | -1,19 | | | | | | | |

Itération 3

| | Univ | Cpge | Autres | Pond | V4RPnd | V6CNP | **V6CPnd** | CarPnd |
|---|---|---|---|---|---|---|---|---|
| L | 0,04 | 0,00 | -0,04 | 0,20 | 0,81 | 0,08 | **0,42** | 0,04 |
| ES | 0,06 | -0,01 | -0,05 | 0,20 | 1,09 | 0,11 | **0,57** | 0,06 |
| S | 0,00 | 0,02 | -0,02 | 0,30 | 0,17 | 0,03 | **0,09** | 0,00 |
| Tech | -0,10 | -0,01 | 0,11 | 0,30 | -1,44 | -0,23 | **-0,75** | 0,17 |
| Pond | 0,50 | 0,10 | 0,40 | | | | Somme | 0,2722 |
| V5CNP | 0,24 | 0,01 | -0,25 | | | | Norme | 0,5217 |
| V5CPnd | 0,48 | 0,07 | -0,62 | Somme | Norme | | ***Stop*** | |
| CarPnd | 0,12 | 0,00 | 0,15 | 0,2722 | 0,5217 | | | |
| V5RNP | 0,46 | 0,01 | -0,48 | | | | | |
| V5RPnd | 0,93 | 0,13 | -1,19 | | | | | |

Recherche des vecteurs propres en analyse factorielle des correspondances. La convergence de l'algorithme est obtenue car les vecteurs V4 et V6 calibrés pondérés sont identiques

| | Univ | Cpge | Autres | Pond | V6CPnd |
|---|---|---|---|---|---|
| L | 0,039 | 0,030 | -0,264 | 0,20 | 0,42 |
| ES | 0,275 | 0,040 | -0,354 | 0,20 | 0,57 |
| S | 0,043 | 0,006 | -0,055 | 0,30 | 0,09 |
| Tech | -0,363 | -0,052 | 0,467 | 0,30 | -0,75 |
| Pond | 0,50 | 0,10 | 0,40 | | |
| V5CPnd | 0,48 | 0,07 | -0,62 | Norme= | 0,52 |

Reconstitution de l'approximation du premier facteur
Par ex. pour la case L Univ : 0,039 = 0,42 x 0,48 x 0,2 x 0,5 / 0,52

Le reste se déduit en soustrayant du tableau des écarts le tableau d'approximation. Pour obtenir les vecteurs propres du deuxième facteur, il faut recommencer l'algorithme d'extraction.

*Les contributions en analyse des correspondances*

Dans le tableau ci-dessous, on a mis en colonne gauche le tableau des écarts à l'indépendance et sa décomposition en deux facteurs. Dans le premier tableau de droite, on a calculé de manière tout à fait traditionnelle le khi-deux de chaque case sur le tableau en proportion (ce qu'on appelle dans ce cas le phi-deux). Par exemple pour la première case, L x Univ l'écart est de 0,040 et son carré de 0,0016 ; l'effectif théorique est le produit des marges 0,2 x 0,5 = 0,1 ; la contribution de la case (écart$^2$ / théorique) est donc de 0,0016 / 0,1 = 0,016. On effectue les totaux en ligne et en colonne.

Tableau des écarts

| | Univ | Cpge | Autres | Pond |
|---|---|---|---|---|
| L | 0,040 | 0,000 | -0,040 | 0,20 |
| ES | 0,060 | -0,010 | -0,050 | 0,20 |
| S | 0,000 | 0,020 | -0,020 | 0,30 |
| Tech | -0,100 | -0,010 | 0,110 | 0,30 |
| Pond | 0,50 | 0,10 | 0,40 | |

Khi-deux des écarts

| | Univ | Cpge | Autres | Total | | |
|---|---|---|---|---|---|---|
| L | 0,016 | 0,000 | 0,020 | 0,04 | | |
| ES | 0,036 | 0,005 | 0,031 | 0,07 | | |
| S | 0,000 | 0,013 | 0,003 | 0,02 | | |
| Tech | 0,067 | 0,003 | 0,101 | 0,17 | | **%** |
| Total | 0,12 | 0,02 | 0,16 | **0,2958** | | **100** |

Premier facteur

| | Univ | Cpge | Autres | Pond |
|---|---|---|---|---|
| L | 0,039 | 0,001 | -0,040 | 0,20 |
| ES | 0,053 | 0,002 | -0,054 | 0,20 |
| S | 0,012 | 0,000 | -0,013 | 0,30 |
| Tech | -0,104 | -0,003 | 0,107 | 0,30 |
| Pond | 0,50 | 0,10 | 0,40 | |

Khi-deux du premier facteur

| | Univ | Cpge | Autres | Total | % | |
|---|---|---|---|---|---|---|
| L | 0,015 | 0,000 | 0,020 | 0,04 | 13,2 | |
| ES | 0,028 | 0,000 | 0,037 | 0,06 | 23,8 | |
| S | 0,001 | 0,000 | 0,001 | 0,00 | 0,9 | |
| Tech | 0,073 | 0,000 | 0,096 | 0,17 | 62,1 | |
| Total | 0,12 | 0,00 | 0,15 | **0,2722** | 100 | **92,0** |
| % | 43,0 | 0,2 | 56,9 | 100 | | |

Deuxième facteur

| | Univ | Cpge | Autres | Pond |
|---|---|---|---|---|
| L | 0,001 | -0,001 | 0,000 | 0,20 |
| ES | 0,007 | -0,012 | 0,004 | 0,20 |
| S | -0,012 | 0,020 | -0,007 | 0,30 |
| Tech | 0,004 | -0,007 | 0,003 | 0,30 |
| Pond | 0,50 | 0,10 | 0,40 | |

Khi-deux du deuxième facteur

| | Univ | Cpge | Autres | Total | % | |
|---|---|---|---|---|---|---|
| L | 0,000 | 0,000 | 0,000 | 0,00 | 0,3 | |
| ES | 0,001 | 0,007 | 0,000 | 0,01 | 31,3 | |
| S | 0,001 | 0,013 | 0,000 | 0,01 | 60,7 | |
| Tech | 0,000 | 0,002 | 0,000 | 0,00 | 7,7 | |
| Total | 0,00 | 0,02 | 0,00 | **0,0236** | 100 | **8,0** |
| % | 7,0 | 89,8 | 3,1 | 100 | | |

Contributions en analyse des correspondances

La somme des contributions est de 0,2958 (en terme de khi-deux il faudrait multiplier par l'effectif, ici c'est un phi-deux. Dans la suite et pour faciliter la compréhension, je parlerai de khi-deux d'un tableau en proportion, identique de ce fait au phi-deux).

Les écarts se décomposent en deux fragments, ceux du premier facteur puis ceux du deuxième. Si l'on ne prenait en compte que le premier facteur, on pourrait d'une manière analogue calculer les contributions au khi-deux de chaque case. Par exemple pour la première, au lieu de prendre un écart de 0,040 on prendrait son approximation 0;039. En faisant de même pour toutes les cases, avec toujours les mêmes effectifs théoriques, on a dans le tableau de droite les contributions au khi-deux du premier facteur. On fait de même avec les écarts restant du deuxième facteur. On voit alors que :

1) le khi-deux de départ est strictement égal au khi-deux du premier facteur ajouté au khi-deux du deuxième facteur,

2) que cette répartition est très inégalitaire : la plus grande partie du khi-deux, indicateur de l'information apportée, se trouve dans le premier facteur. Comme la distribution entre les deux facteurs est additive (0,2722 + 0,0236 = 0,2958) on peut regarder en pourcentage l'apport de chaque facteur. Le premier apporte 92% du total, le deuxième 8%. Le premier facteur est la bonne approximation, le deuxième n'est qu'un reste éventuellement négligeable. Ces pourcentages sont appelés aussi taux d'explication.

3) ces sommes des khi-deux de chaque facteur correspondent aux valeurs propres de la décomposition factorielle.

4) les totaux dont ils sont issus sont eux-mêmes les sommes des carrés pondérés des vecteurs propres calibrés, ce qu'on peut vérifier ici sur le premier facteur où les totaux en ligne et en colonne correspondent aux carrés pondérés (CarPnd) des vecteurs V5 en ligne et V6 en colonne.

5) Ces totaux sont appelés contribution absolue du facteur. Elles peuvent être mises en rapport avec le total, c'est-à-dire la valeur propre de chaque facteur. On voit que la contribution relative du premier facteur la plus forte est issue du bac technique. Avec les contributions on peut ainsi voir ce qui a fait un facteur, ce qui en facilite l'interprétation

6) La décomposition se fait aussi en prenant en compte chaque total de ligne (ou colonne) : le total du tableau d'origine est aussi égal à la somme des totaux des deux facteurs. Par exemple ici le total du khi-deux de la ligne ES 0,07 est égal au khi-deux de ES pour le premier facteur 0,06 + celui du deuxième facteur 0,01. En pourcentage ou en proportion, cet indicateur est souvent appelé "cosinus²" car il peut être interprété comme tel. Pouvant être redondant avec la contribution par facteur, il n'est pas donné dans tous les programmes.

*La présentation logicielle standard*

Dans trideux, les résultats sont les suivants:

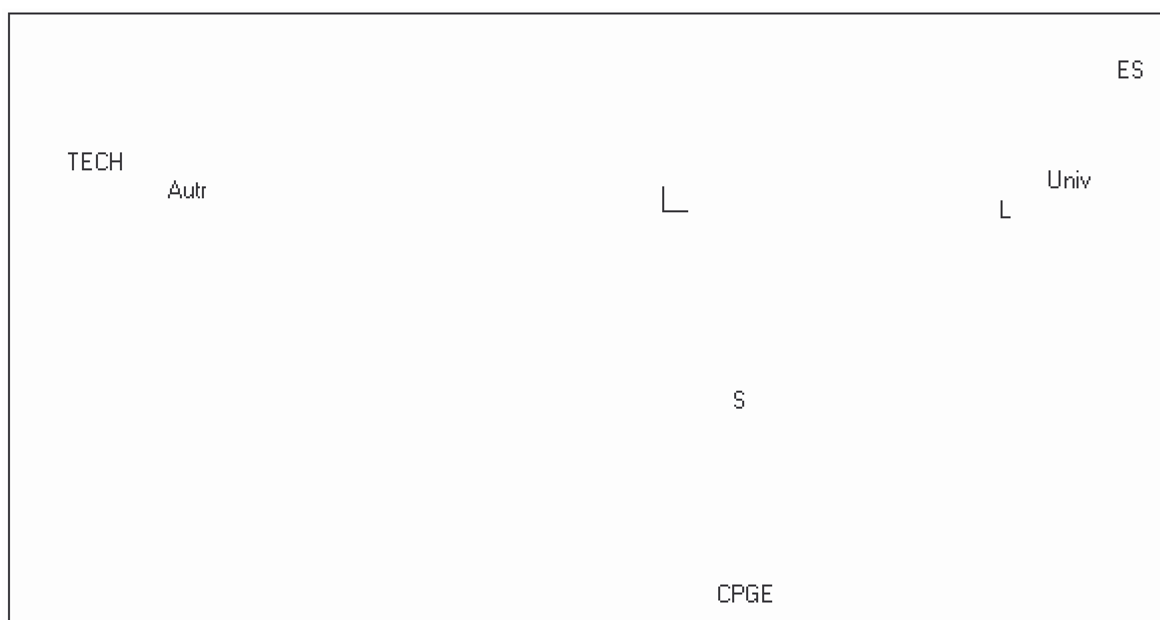
```
Le phi-deux est de :      0.295750
Facteur   1 Valeur propre =  0.272152 Pourcentage du total = 92.0
Facteur   2 Valeur propre =  0.023598 Pourcentage du total =  8.0
Coordonnees factorielles (F= ) et  contributions pour le facteur (CPF)
*---*------*----*------*----* Lignes du tableau
ACT.    F=1  CPF    F=2  CPF
*---*------*----*------*----*
L       424  132     19    3
ES      570  238    192  312
S        88    9   -219  607
TECH   -751  621     78   78
*---*------*----*------*----*
*   *      *1000*      *1000*
*---*------*----*------*----* Modalites en colonne
ACT.    F=1  CPF    F=2  CPF
*---*------*----*------*----*
Univ    484  430     57   70
CPGE     69    2   -460  898
Autr   -622  568     43   32
*---*------*----*------*----*
*   *      *1000*      *1000*
*---*------*----*------*----*
```

Après l'indication de la décroissance des valeurs propres en phi-deux et en pourcentage, les lignes et colonnes sont présentées de la même façon. Sous la rubrique F1 on trouve les coordonnées du vecteur propre (calibré, pondéré et en millièmes), puis sous la rubrique CPF la contribution au facteur du khi-deux de chaque ligne ou colonne.

Pour l'interprétation, on se sert à la fois du signe de chaque coordonnée du vecteur propre et de sa contribution. On voit que le premier facteur est fait d'une opposition entre l'enseignement technique (côté négatif, donc à gauche sur le graphique) qui s'oppose aux séries L et ES. Cette opposition ne prend pas en compte le bac S qui n'apporte que 9 pour mille de contribution. De même pour les colonnes, l'université (côté positif) s'oppose aux autres orientations et les classes préparatoires ne sont pas prises en compte par ce facteur.

Pour le deuxième facteur (qui n'apporte globalement que 8% de l'information), l'opposition se crée maintenant entre S (60,7%) côté négatif, donc en bas sur le graphique, qui s'oppose à ES en haut (31,2%) Technique et surtout L sont peu pris en compte par ce facteur. Pour les colonnes, CPGE en bas s'oppose au reste.

On peut donc dire que le premier facteur oppose l'enseignement technique à l'enseignement général et que le deuxième distingue dans cet enseignement général entre la filière université et la filière grandes écoles. Ce qu'indique aussi le plan factoriel du premier facteur horizontal et du deuxième vertical.




Analyse des correspondances : plan du premier et deuxième facteur